The Grammar of Transformers: A Systematic Review of Interpretability Research on Syntactic Knowledge in Language Models

Anonymous ACL submission

Abstract

We present a systematic review of 337 articles evaluating the syntactic abilities of Transformer-based language models (TLMs), reporting on over 3,000 datapoints spanning a wide range of syntactic phenomena, languages, models, and methods. We take the data to collectively show that TLMs encode a non-trivial amount of syntactic knowledge. Behavioral evidence shows strong performance on formal syntactic phenomena, but weaker and more variable performance on phenomena at the syntax-semantics interface. Performance is also consistently lower for languages with less digital support. Probing and mechanistic studies further support the presence of syntactic knowledge in TLMs. Yet, because most work remains observational and current approaches are methodologically heterogeneous, insight into the detailed computational mechanisms underlying syntactic processing remains limited. At the same time, the literature remains heavily concentrated on English and BERT-like models. We discuss the implications of our results and provide recommendations for future research.

1 Introduction

It has been observed at least since Gulordava et al. (2018) that language models (LMs), though trained simply on a token prediction task, seem to acquire a non-trivial amount of syntactic knowledge. In recent years, a growing body of research has focused on understanding what precisely is learned and how it is used at processing time (e.g., Rogers et al., 2020; Lasri et al., 2022a; Brinkmann et al., 2025). However, even if some trends have emerged, there is yet no general picture about syntactic knowledge in LMs. We therefore think that it is a good moment to stop and appraise the extant literature. In this article, we do precisely that, through a **systematic review** (Page et al., 2021), a type of review that (to our knowledge) is not commonly used in computational linguistics (CL) and NLP. Systematic reviews differ from common narrative reviews in that they include transparent search, selection, and annotation criteria, and they provide quantitative evidence for the identified patterns. We focus on Transformer-based language models (TLMs) because they are the dominant model as of 2026.

The specific goals for this review are twofold: (i) mapping the landscape of research on syntactic knowledge in TLMs (what has been done and how; what remains to be done), and (ii) summarizing what is currently known about syntactic knowledge in TLMs.

Our work builds on six previous reviews with overlapping content: Limisiewicz and Mareček (2020), Linzen and Baroni (2021), Kulmizev and Nivre (2022), Chang and Bergen (2023), Millière (2024), and López-Otal et al. (2025). These reviews found that modern LMs acquire rich syntactic representations early during pretraining (Chang and Bergen, 2023), localize much of this information in their middle layers, and encode at least some syntactic relations via specialized attention heads and groups of neurons (López-Otal et al., 2025). The reviews also point out limitations of TLMs, including reliance on surface-level heuristics, lexical sensitivity, and challenges with rare constructions and subtle distinctions (Kulmizev and Nivre, 2022; Linzen and Baroni, 2021). While valuable, these prior reviews are largely narrative and interpretive. In contrast, we provide a quantitative and systematic synthesis of the literature with the aim of offering a transparent mapping between the available evidence and the conclusions, an approach that is increasingly necessary given the scale and methodological heterogeneity of the knowledge accumulated in recent years.

Our review covers 337 articles and 3,074 individual datapoints across different languages, syn-

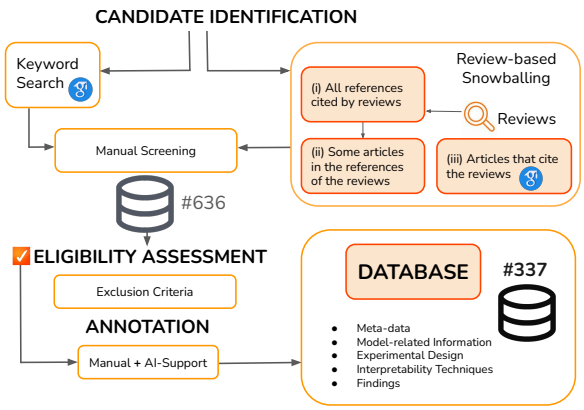


Figure 1: Flowchart summarizing database creation.

tactic phenomena, methods and datasets. We also contribute to the community an annotated database of the reviewed studies, to be made publicly available alongside the analysis code upon acceptance.

2 Method

Fig. 1 summarizes the process we followed to create the database, comprising paper candidate identification, eligibility assessment, and annotation. We did data collection in summer 2025, with a cut-off date for inclusion of July 31, 2025. Importantly, the database is designed to remain open to future extensions, with a form[1] allowing researchers to contribute and add new papers.

Paper candidate identification. In the initial identification stage, we prioritized recall, deferring concerns about precision to the subsequent stage. Article relevance was primarily assessed via its title (in case of doubt, we reviewed the abstract, and, if necessary, the full text). To identify candidates, we followed a two-pronged strategy: A *keyword search* in Google Scholar, using keywords that combined "syntactic structure/knowledge" and "LLMs/LMs/transformer", and a *snowballing strategy* using the six aforementioned reviews on the topic to identify relevant references. In the latter, we screened (i) all the references cited by the reviews,[2] resulting in 263 articles, (ii) some of the articles that were cited in these 263 articles, and (iii) most of the articles that cite the reviews themselves, identified via Google Scholar.[3] This yielded 636 articles.

Eligibility assessment. This stage shifted attention to precision. We excluded: 1) Studies focusing on non-transformer architectures (e.g., RNNs, LSTMs) or on encoder–decoder architectures, particularly those designed for machine translation, so as to keep articles within the scope of our research question. 2) Studies that, while analysing TLMs, do not empirically assess syntactic knowledge. 3) Publications without a clear empirical component, such as position papers or surveys (the previous reviews were not included in the analysis due to the absence of an empirical component). 4) Publications that were neither peer-reviewed journal articles, conference papers, nor archived preprints. This implies excluding blog posts and MA or PhD theses, since they often report on findings that are also published in peer-reviewed venues. The remaining 337 articles (cited in App. A, Table 1) were further annotated and analysed .[4]

Annotation. We defined 33 annotation variables, organized into five categories, to systematically capture the aspects of the selected works that would enable quantitative analyses and cross-study comparisons. An overview is provided below, with a detailed description in App. B.

The *meta-data* category includes publication details and information on how articles entered or exited the review process (source of each article and, if applicable, the grounds for exclusion). *Model-related* information specifies the model name(s), type(s), fine-tuning status, and language coverage. *Experimental design* codes the syntactic phenomena investigated, languages tested, and experimental materials employed. To establish the syntactic categories, we applied a bottom-up approach, starting from the terms used in the studies themselves and grouping them under broader labels recognizable to both the linguistics and NLP communities. The resulting 11 labels are shown in Fig. 2 (right) and described in detail in App. B. *Interpretability techniques* include the specific method, as named in the original paper, and a general category (behavioral, mechanistic, or probing, along the lines of Millière, 2024, see next section for more information). Further, we annotated various dimensions for probing and mech-

[1] Link after acceptance.

[2] Except for López-Otal et al. (2025), who provided a dedicated section of syntax-related studies in TLMs, and where we only considered the studies listed in that section.

[3] If a review had more than 100 citations, we reviewed the top 30-40% (sorted by relevance as determined by Google Scholar), otherwise all citations were checked.

[4] Excluded articles and exclusion reasons were retained for transparency and future studies beyond the scope of this review.

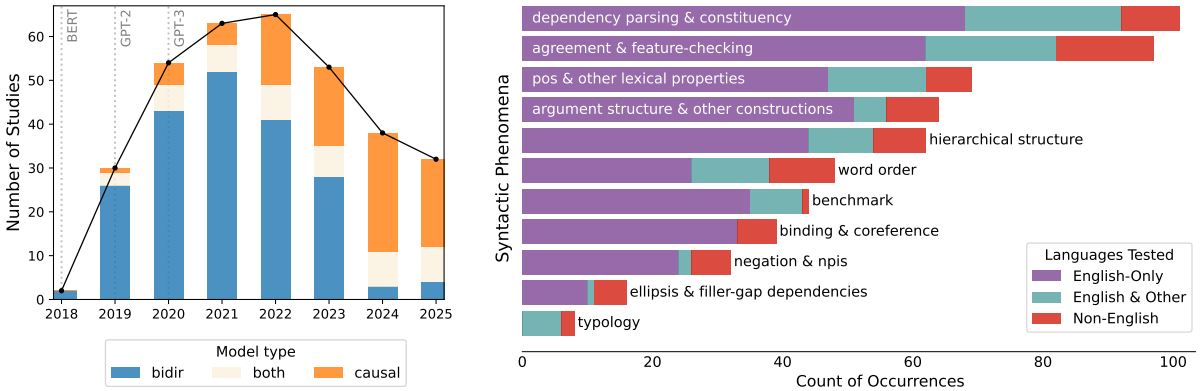


Figure 2: **Left:** Number of studies published each year; debut of influential models. **Right:** Counts of syntactic phenomena by evaluated language (only English tested, English and other languages, and non-English).

anistic studies, covering methodological aspects, causal nature, locality and strength of evidence of syntactic awareness. The *findings* category summarizes, in prose, the main insights regarding syntactic competence and processing as well as any model comparisons.

We produced annotations manually for 28 of the 33 categories; for the remaining 5 (model type, syntactic phenomena, experimental materials, method name, and main syntactic findings) we used a semi-auomated AI-assisted workflow to extract some information from the papers; final labels were determined manually. The AI workflow was successfully validated against a manual gold standard (further details are provided in App. C).

3 A Quantitative Overview of the Field

Fig. 2 (left), which shows the distribution of papers by year, indicates that interest in the topic of our review quickly took off after the appearance of BERT (Devlin et al., 2019), peaking in 2022 and decreasing afterwards (though note that the 2025 data only covers up to July, so the final count will likely exceed that of 2024). As expected, initially most work focused on bidirectional TLMs, but in later years, causal models have dominated. Also note that most studies focus on one model type (either bidirectional, 59%, or causal, 28%), with very few comparing the two (13%).

Beyond the information shown in Fig. 2 (left), we found that studies predominantly evaluate monolingual models (66%, the vast majority English), rather than multilingual (20%) or both kinds (14%). Also, consistent with previous observations (Waldis et al., 2024; López-Otal et al., 2025), most of the results in our database concern a single model: BERT, with 30% of the results. BERT plus variants account for 58% of the datapoints, and the four overall most studied models for 62%. The over-focus on BERT is likely due to BERT's early availability, strong baseline performance, and suitability for a wide range of research questions. As the field matures, research is increasingly moving toward a broader and more principled selection of architectures.

Linguistic phenomena & language diversity. Fig. 2 (right) shows that the distribution of syntactic phenomena is severely imbalanced: Most research has been conducted on formal syntax (c.f. *dependency parsing and constituency*, *agreement and feature-checking*, *POS and other lexical properties*—with most studies in this category focusing on POS—, *hierarchical structure*, and *word order*). These are phenomena that are relatively easy to get at with current NLP tools, especially for English. Instead, phenomena at the syntax-semantics interface have received much less attention (cf. *binding and coreference*, *negation and NPIs*, *ellipsis and filler–gap dependencies*); and only a handful of studies concern typological aspects. Somewhat surprisingly, only 42 studies used benchmarks, 29 of which used BLiMP (Warstadt et al., 2020a).[5] Fig. 2 (right) also showcases the overwhelming dominance of English (in purple), which we already noted regarding the models (here, the data shows the languages *tested* in the articles, not the languages covered by the model).

Methods. Recall from above that we follow Millière (2024)'s classification of interpretabil-

[5]The others were the CoLA task from GLUE (Wang et al., 2018, 11 uses), SyntaxGym (Gauthier et al., 2020, 3 uses), Holmes and HANS (McCoy et al., 2019; Waldis et al., 2024, 1 use each).

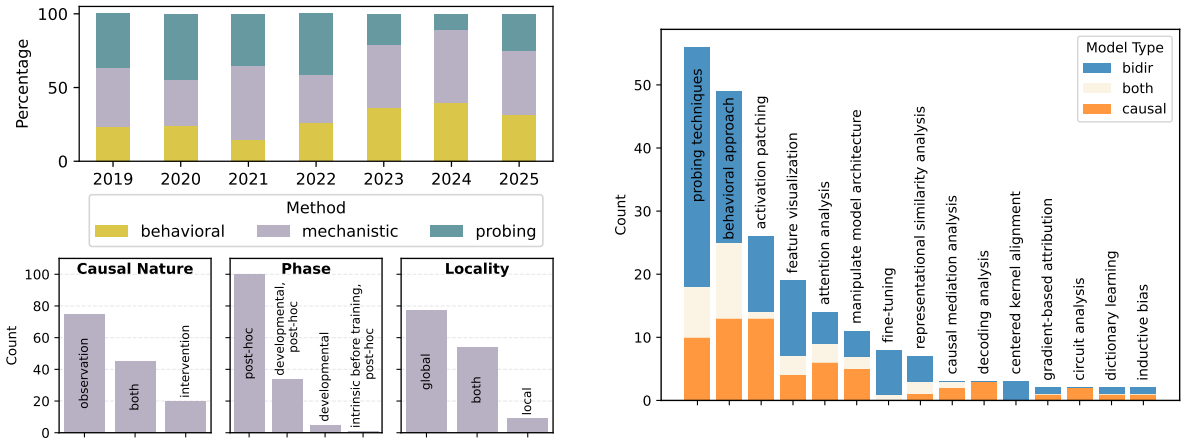


Figure 3: **Left:** Top: Percentage of methods over years. Bottom: mechanistic interpretability methods broken down by causal nature, training phase, and locality. **Right**: Most frequent interpretability techniques used in mechanistic papers; methods may be combined.

ity methods into behavioral, probing, and mechanistic, as it provides a useful framework for organizing the literature. Fig. 3 (left-top) shows a relatively balanced distribution of these methods over the years. Behavioral methods assess models based on their output behavior at the final layer. They include analysis of generated outputs and comparisons of model probabilities in controlled settings (e.g., minimal pairs, Linzen et al., 2016; Lau et al., 2017). The latter paradigm, highly influenced by the psycholinguistic tradition, treats the higher probability of a correct form compared to an incorrect alternative as evidence that the TLM has mastered a given syntactic construct. Probing is a specific interpretability method that became very popular after Hewitt and Manning (2019) introduced it, which is why it is in a category on its own. It aims at detecting specific types of linguistic information in the models' internal representations, most often by training auxiliary classifiers on activations from specific layers (Waldis et al., 2024). The argument goes that if, for instance, a classifier trained to recover the POS of input words from a given hidden layer is successful, then that layer encodes POS information. Finally, mechanistic approaches actually cover a wide array of specific methods, see Fig. 3 (left-top) for the most common ones. These methods may (and often do) incorporate behavioral or probing elements. In our database, we categorized studies investigating internal dynamics, even if including behavioral or probing tests as well, as mechanistic. Overall, mechanistic approaches aim at getting at internal mechanisms by inspecting model components (e.g., layers, neurons, circuits, or training dynamics) and, where possible, intervening on them to causally link representations to behavior (Wiegreffe and Saphra, 2024). For instance, these methods may modify the activations of components identified as syntactically crucial and test the effects of these interventions on model behavior.

All methods have their pros and cons. For instance, probing offers scalability and clear feature detection but, on its own, lacks causal insight: decoding linguistic features is not equivalent to showing that the identified representations are actually used in predictions (besides other issues, see Belinkov, 2022; Wiegreffe and Saphra, 2024; Agarwal et al., 2025). Therefore, it is reassuring that the field has been using a healthy variety of methods throughout (Fig. 3, left-top). The only significant temporal trend is that probing methods have diminished substantially in recent years, possibly due to the growing awareness of their limitations.However, there is one aspect that we believe should improve: There are still too few studies that can get at causal mechanisms (only 19% use interventional methods, see App. D, Fig. 13).

We next delve a bit more into mechanistic studies, largely following the taxonomy from Bereska and Gavves (2024). We annotated the 140 mechanistic papers in our database by considering (i) the causal nature of the specific method (observational, interventional, or both; Fig. 3, left-bottom, *Causal Nature*); (ii) which stage of the TLM lifecycle is investigated, intrinsic (before training), developmental (during training), or post-hoc (after training; *Phase* in Fig. 3); (iii) which part or scale of the model the analysis targets, ranging from global representations (e.g., the residual stream)

to local units (e.g., individual neurons; *Locality*); and (iv) the specific methods used, as described by the authors (with some generalization; the most frequent are shown in Fig. 3, right). We find over 80 specific methods in mechanistic studies, which we take to be a positive development since interpretability is still young field.[6] Post-hoc analysis of fully trained models at a global level clearly dominates. However, there is a healthy amount of papers that track development, and many also check both global and local trends. Additional results for this section can be found in App. D.

4 Syntactic Knowledge in TLMs

We report on behavioral evidence derived from papers that measure model accuracy, before turning to results from probing and mechanistic approaches. We provide additional results for this section, including complete statistical model specification, in App. E.

4.1 Behavioral Evaluation

In the following, we report on benchmark results for English (as this is a controlled setup where results are comparable) and on results from non-benchmark studies for 113 languages (offering a broader but more heterogeneous view).

BLiMP. The BLiMP benchmark (Warstadt et al., 2020a) is designed for behavioral evaluation of models through minimal pair comparison. Our analysis includes English BLiMP scores from 25 papers: 11 that report by-category results and 14 that report average BLiMP scores. Fig. 4 (left) shows accuracies by phenomenon (an average of 31 datapoints per category, as most articles analyze multiple models; note that we include only results from fully trained models). The most salient feature of these results is how well the models perform on this benchmark. Notably, the best TLM is at or above human performance for 7 out of the 11 categories, and all but 8 out of the 368 results are above the baseline. Another salient feature is the large within-phenomenon variance across TLMs. We find that this variance is mostly driven by how good each model is at syntax generally, as measured by its overall BLiMP score: Model rankings across phenomena are highly consistent. This is illustrated in Fig. 4 (left) by highlighting the best (purple star) and worst (yellow dot) TLMs, and Fig. 16 in App. E.1 provides quantitative support. In this respect, additional analyses of average BLiMP scores confirm the expected scaling trend: larger TLMs and TLMs trained on more data generally achieve higher performance (see App. E.1).

Despite the large variance, the results also clearly suggests that TLMs perform better on formal syntactic relations (e.g. all agreement categories have medians over 85%) than on phenomena at the syntax-semantics interface (binding, argument structure, NPI licensing, control/raising, quantifiers and island effects all have medians below 75%). One exception is ellipsis: despite its reliance on interactions between syntax and semantics, several TLMs match or outperform humans on this category. However, BLiMP evaluates only specific cases of noun phrase ellipsis, whose violations may be comparatively easy to detect (Warstadt et al., 2020a), suggesting this result may reflect dataset characteristics rather than ellipsis as a phenomenon being easy for models.

More generally, despite its value for enabling cross-model comparisons across phenomena, BLiMP still provides limited coverage of syntax, even for English. It focuses on cases that can be readily expressed as minimal pairs and aggregates heterogeneous cases into composite scores with equal weighting, potentially obscuring important differences in difficulty and relevance (Kulmizev and Nivre, 2022). As a result, model performance should be interpreted with the characteristics and limitations of the benchmark in mind.

Beyond benchmarks and English. We next aggregate all the non-benchmark accuracy scores from the database, yielding 1,557 datapoints from 76 studies spanning 113 languages. Fig. 4 (right) presents results for our annotated categories of linguistic phenomena. The variance this time is even greater than in the BLiMP case, which is not surprising given that different studies use different materials and test different languages. However, since this makes for an unclear picture at a descriptive level, we turn to statistical modeling to investigate the factors influencing accuracy. In particular, we use as predictors syntactic phenomena, model type (causal / bidirectional), and language support to see how syntactic knowledge

[6] Recall that mechanistic papers often include also probing and behavioral methods, and indeed they show in the figure as the two most frequent approaches even within mechanistic papers.

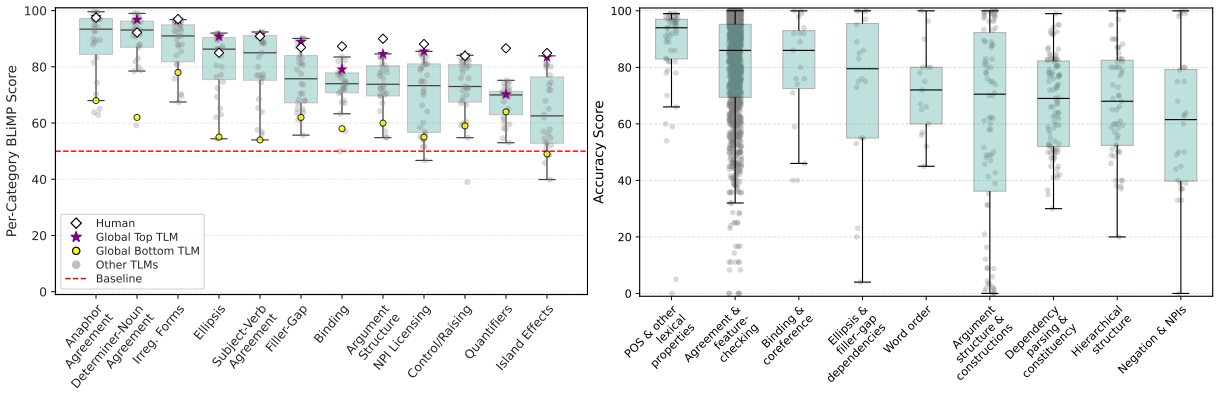


Figure 4: **Left**: **BLiMP (English-Only)** scores by category. Human performance ◇ from Warstadt et al. (2020a). Top ⋆ (Roberta large) and bottom ● (Minibert) TLM among those with complete per-category results. **Right**: **Accuracy scores (113 languages)** by linguistic phenomena, covering more than 1,500 data points from 76 studies. Note that 77% of the datapoints for Agreement & feature-checking originate from one study (Jumelet et al., 2025).

varies across languages. Language support gets at the availability of NLP resources and datasets for a given language. We take the support data from Joshi et al. (2020), who define a 5-level classification ranging from low- to high-support. We fit a regression model and find that both model type and language support are significant predictors of accuracy scores: causal models outperform bidirectional models by 10% points ($\beta = 0.1$, $p < 0.001$). In addition, as could be expected, performance increases with language support, with each higher support category associated with a gain of 5% points ($\beta = 0.05$, $p < 0.001$). The estimated intercepts for the linguistic phenomena categories (App. E.2, Fig. 18) confirm that *POS and other lexical properties* and *agreement and feature-checking* have generally higher accuracies than average, while *hierarchical structure*, *negation & NPIs* and *argument structure & constructions* have lower accuracies. For the other phenomena, there is no evidence either way. Thus, this analysis largely confirms that formal syntactic phenomena are easier for TLMs than phenomena at the syntax-semantics interface.

4.2 From Behavior to Internal Mechanisms

Probing. Probing has been used to localize layers where syntactic information is most readily available. We code the the 56 probing-based studies that include this information (including mechanistic papers that also use probing) for the model locations at which studies report the strongest probing performance. Fig. 5 (left) shows that the middle layers are most commonly identified, corroborating previous observations (López-Otal et al., 2025).[7] This aligns with a view of TLM representations as distributed across layers, motivating analyses of how information accumulates and transforms throughout the network. In this vein, some mechanistic studies examine progressive knowledge encoding (e.g., He et al., 2025b), consistent with staged interpretations of decoder-only models that reflect a transition from surface-level cues to deeper syntactic and hierarchical processing (Lad et al., 2024).[8]

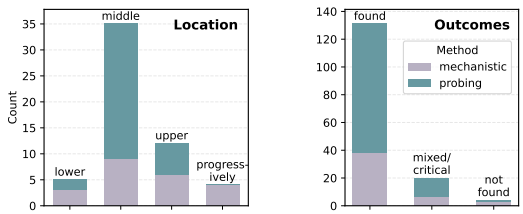


Figure 5: Probing-based studies. **Left:** location in TLMs where probing performance/syntactic knowledge is strongest. **Right:** author-reported evidence for syntactic representations in TLMs. Note that some mechanistic papers include probing methods; these are included.

Furthermore, we check 147 probing-based studies for their outcomes: Each study was coded according to whether probing recovered syntactic information from TLM representations, yielded mixed results, or failed to recover such infor-

[7] Despite the prominence of middle layers, studies typically describe them as showing stronger, rather than exclusive, encoding of syntactic information. We also observed that the distinction between "middle" and "upper" layers is somewhat subjective and varies across studies.

[8] Unlike a strict pipeline interpretation (e.g., Jawahar et al., 2019), this staged view does not assume discrete functional modules, but rather a gradual and overlapping transformation of representations across depth (e.g., Niu et al., 2022b).

mation. Fig. 5 (right) shows the results. 85% of studies report successful recovery of syntactic information from model representations. Although this high success rate may partly reflect publication bias toward positive results, this again points to substantial syntactic knowledge in TLMs. Additional analyses do not reveal substantial differences across syntactic phenomena in whether studies successfully recover syntactic information from model representations (App. E.3, Fig. 20). This contrasts with behavioral results, which do show clear differences in how easily models handle different syntactic phenomena. The mismatch provides further support to the idea that successful probing is not, by itself, a reliable measure of the quality or functional use of syntactic representations in TLMs (Agarwal et al., 2025).

Mechanistic interpretability. We tried to systematize the mechanistic studies to obtain more insights into the syntactic mechanisms in TLMs, but we faced a major challenge: once studies are broken down by method and phenomenon to enable meaningful comparisons, the resulting groups become too sparse and heterogeneous for a straightforward synthesis. We therefore decided to instead focus on author-reported strength of evidence for the presence of syntactic knowledge, similarly to what we did for probing studies. The results, shown in Fig. 6, are positive: 79% of the 140 mechanistic papers in our database find clear supporting evidence. At the same time, Fig. 6 again underscores that the vast majority of this evidence comes from studies focusing exclusively on English. Again, we fail to observe salient differences across phenomena, but all the syntax-semantics-interface phenomena have low coverage in the mechanistic literature, so this is still an open point (see App. E.4, Fig. 21).

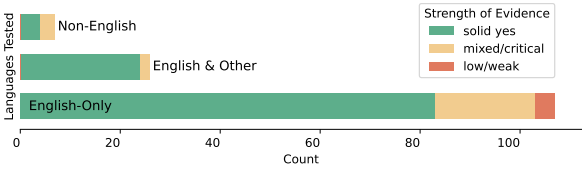


Figure 6: Evaluation of syntactic awareness in mechanistic interpretability studies by language tested.

5 Discussion & Conclusion

Conducting systematic reviews in CL/NLP. To our knowledge, this is among the first systematic reviews in CL and NLP. As the field of AI continues to grow at breakneck speed, systematic reviews and meta-reviews will become ever more important, as in other large fields like Medicine. However, conducting such reviews is challenging, and we have found some current practices in the community that make it even more challenging. We propose two main recommendations moving forward. First, studies should report complete data, to allow for meaningful comparison of results. Many papers included only partial data on models and methods, despite the fact that results crucially depend on factors like architecture, size, tokenization, and training data. Benchmark reporting is also often partial (e.g., only 11/25 papers include BLiMP per-category results). Additionally, documenting task framing and prompting is crucial, as these can also significantly impact outcomes (e.g., zero-shot vs. few-shot). Second, we should collectively standardize investigations to a greater extent. Differences in how syntactic phenomena are defined fragment results and limit comparability, as even the notion of "syntactic knowledge" varies across studies.[9] We offer a typology of syntactic phenomena that has been built in a bottom-up fashion from the current literature, and which could be refined and expanded upon as the field grows.

Standardization is particularly important for cross-linguistic research. We have shown a huge variance in cross-linguistic performance, only partially accounted for factors like the digital support of a language or the type of linguistic phenomenon. One of the main drivers, presumably, are the characteristics of the specific data the models are being tested on, but those can vary; compare for instance determiner-noun number agreement (easy) versus long-distance subject-object agreement (hard). Therefore, a standardization effort needs to be made in creating multilingual benchmarks, beyond direct translations or adaptations of English datasets (e.g., as in Baucells et al., 2025; Jumelet et al., 2025), as we otherwise risk overlooking language-specific phenomena. As an example of a "best of both worlds" approach that both standardizes and allows for variation, the Spanish variant of SyntaxGym (Pérez-Mayos et al., 2021) incorporates language-specific

[9] For instance, Waldis et al. (2024) place several of the phenomena we have considered here outside of syntax (e.g. negation in reasoning, and negative polarity item licensing in semantics).

features like subjunctive mood and flexible word order besides English-compatible phenomena.

The state of the field. Our review reveals a vibrant area of research, with plenty of studies and a healthy variety of methods (337 papers and more than 3,000 individual results are included in our database). The literature also shows substantial transfer of theoretical and methodological insights from psycholinguistics to model interpretability, enabling more principled comparisons between human and model behavior and strengthening links to linguistic theory.

However, on the negative side, we find an overstudy of a single language (69% of the articles are on English only, and 91% include English) and a few models (30% of the results concern BERT; including BERT and its variants raises the total to 58%), which severely limits the robustness and generality of the conclusions. In addition, while formal syntactic phenomena such as agreement, syntactic structure, word order, and morphosyntactic properties such as part of speech have been well studied, the syntax–semantics interface remains comparatively underexplored: phenomena such as binding, coreference, negation, ellipsis, or filler-gap dependencies have received very little attention. Expanding research to cover the syntax-semantics interface is particularly necessary given that we have found that in general models perform worse at it than at more formal aspects of syntax.

At a methodological level, at this stage of the research, some methods are established (behavioral, probing) and others are still emerging (under the broad umbrella "mechanistic"). Overall, however, we find an overuse of purely observational methods (81%), at the expense of causal approaches that can link model behavior to internal mechanisms. While we encourage the use of interventional methods, we emphasize that no single method is sufficient in isolation, and that the strongest evidence ultimately comes from combining observational and interventional paradigms (e.g., Simon et al., 2024; Agarwal et al., 2025; Acevedo et al., 2026). Also, the field has mostly focused on analyzing trained models at a global level; it should pay more attention to the emergence of syntactic knowledge during training and to the connections between local and global representations and mechanisms.

Syntactic knowledge in TLMs As for what TLMs learn about syntax through the language modeling task, the evidence suggests: a whole lot. For instance, TLMs perform quite well on the English BLiMP benchmark (average score of 72%, against a 50% random baseline), and between 79 and 85% of the analyzed studies identify syntactic knowledge in TLMs. Even accounting for the effect of publication bias (only positive results tend to be submitted), we take the literature to show that TLMs do encode a non-trivial amount of syntactic knowledge.

As for factors influencing performance, as mentioned above, we find that models perform better at formal syntactic phenomena than phenomena close to the syntax–semantics interface. As could be expected, larger models and models trained with more data tend to perform better; and performance degrades for lesser-resourced languages. Importantly, we also find that models that perform well on some syntactic phenomena tend to perform well across others, suggesting a general syntactic ability rather than phenomenon-specific strengths. As for internal mechanisms, probing data supports the view that much of syntactic processing occurs in the middle layers (López-Otal et al., 2025). However, there is still limited mechanistic understanding of why such representations emerge or how they support downstream computation. Given the current heterogeneity of the field, and that mechanistic methodologies and reporting practices are still evolving, we leave a broader synthesis of mechanistic insights to future work.

To sum up, the current literature supports the view *that* TLMs possess remarkable syntactic capabilities. However, understanding *how* this knowledge is acquired and used, as well as how it relates to universality and variation in language, will require more work in both the analysis of existing literature and new studies. We encourage greater methodological harmonization, more explicit theoretical commitments, and a broader empirical landscape. Some open topics are whether TLMs implement near-symbolic mechanisms, as hypothesized Boleda (2025); when and how TLMs fall back on superficial statistical patterns (e.g. performance on agreement, binding, or scope often degrades for rare or nonce lexical items, Lasri et al., 2022a); whether improvements with model scale reflect genuine generalization or massive memorization of specific instances (Lake and Baroni, 2018). We hope that our review, database, and annotation standard contributes to moving the field forward.

Limitations

For feasibility, this study focuses on Transformer-based language models and textual input, thus our findings may not fully generalize to recurrent or other neural architectures. Also, we conducted the keyword search in English, which likely led to under-retrieval of publications written in languages other than English. Note, however, that this was only one of the 3 strategies we used to identify relevant literature. Finally, our database is necessarily incomplete and evolving; the resource is intended to remain interactive and expandable as the field progresses. We see this work as a foundation that invites continued contributions and updates from the research community.

References

Mostafa Abdou, Artur Kulmizev, Felix Hill, Daniel M. Low, and Anders Søgaard. 2019. Higher-order Comparisons of Sentence Encoder Representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5838–5845, Hong Kong, China. Association for Computational Linguistics.

Mostafa Abdou, Vinit Ravishankar, Artur Kulmizev, and Anders Søgaard. 2022. Word Order Does Matter and Shuffled Language Models Know It. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6907–6919, Dublin, Ireland. Association for Computational Linguistics.

Santiago Acevedo, Alessandro Laio, and Marco Baroni. 2026. Differential syntactic and semantic encoding in llms. *Preprint*, arXiv:2601.04765.

Judit Acs, Endre Hamerlik, Roy Schwartz, Noah A. Smith, and Andras Kornai. 2024. Morphosyntactic probing of multilingual BERT models. *Natural Language Engineering*, 30(4):753–792.

Ananth Agarwal, Jasper Jian, Christopher D Manning, and Shikhar Murty. 2025. Mechanisms vs. outcomes: Probing for syntax fails to explain performance on targeted syntactic evaluations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33725–33745, Suzhou, China. Association for Computational Linguistics.

Kabir Ahuja, Vidhisha Balachandran, Madhur Panwar, Tianxing He, Noah A. Smith, Navin Goyal, and Yulia Tsvetkov. 2025. Learning Syntax Without Planting Trees: Understanding Hierarchical Generalization in Transformers. *Transactions of the Association for Computational Linguistics*, 13:121–141.

Ahmed Alajrami and Nikolaos Aletras. 2022. How does the pre-training objective affect what large language models learn about linguistic properties? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 131–147, Dublin, Ireland. Association for Computational Linguistics.

Nura Aljaafari, Danilo S. Carvalho, and André Freitas. 2025. Interpreting token compositionality in LLMs: A robustness analysis. *arXiv preprint*. ArXiv:2410.12924.

Badr AlKhamissi, Greta Tuckute, Yingtian Tang, Taha Osama A Binhuraib, Antoine Bosselut, and Martin Schrimpf. 2025. From language to cognition: How LLMs outgrow the human language network. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24332–24350, Suzhou, China. Association for Computational Linguistics.

Matteo Alleman, Jonathan Mamou, Miguel A Del Rio, Hanlin Tang, Yoon Kim, and SueYeon Chung. 2021. Syntactic Perturbations Reveal Representational Correlates of Hierarchical Phrase Structure in Pretrained Language Models. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 263–276, Online. Association for Computational Linguistics.

Zeyuan Allen-Zhu and Yuanzhi Li. 2024. Interpretability of Language Models for Learning Hierarchical Structures.

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. Probing Linguistic Features of Sentence-Level Representations in Neural Relation Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1534–1545, Online. Association for Computational Linguistics.

Chiara Alzetta, Alessio Miaschi, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2022. Probing BERT on Syntactic Representation Paradigms: Dependency vs Constituency Tree.

Chiara Alzetta, Alessio Miaschi, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2025. Parallel Trees: a novel resource with aligned dependencyand constituency syntactic representations. *Language Resources and Evaluation*.

Ben Ambridge and Liam Blything. 2024. Large language models are better than theoretical linguists at theoretical linguistics. *Theoretical Linguistics*, 50(1-2):33–48.

Afra Amini, Tiago Pimentel, Clara Meister, and Ryan Cotterell. 2023. Naturalistic Causal Probing for Morpho-Syntax. *Transactions of the Association for Computational Linguistics*, 11:384–403.

Tatsuya Aoyama and Nathan Schneider. 2022. Probeless probing of BERT's layer-wise linguistic knowledge with masked word prediction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 195–201, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Tatsuya Aoyama and Ethan Wilcox. 2025. Language Models Grow Less Humanlike beyond Phase Transition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24938–24958, Vienna, Austria. Association for Computational Linguistics.

Suhas Arehalli and Tal Linzen. 2024. Neural Networks as Cognitive Models of the Processing of Syntactic Constraints. *Open Mind: Discoveries in Cognitive Science*, 8:558–614.

Aryaman Arora, Dan Jurafsky, and Christopher Potts. 2024. CausalGym: Benchmarking causal interpretability methods on linguistic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14638–14663, Bangkok, Thailand. Association for Computational Linguistics.

David Arps, Laura Kallmeyer, Younes Samih, and Hassan Sajjad. 2024. Multilingual Nonce Dependency Treebanks: Understanding how Language Models Represent and Process Syntactic Structure. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7822–7844, Mexico City, Mexico. Association for Computational Linguistics.

David Arps, Younes Samih, Laura Kallmeyer, and Hassan Sajjad. 2022. Probing for Constituency Structure in Neural Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6738–6757, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the Cross-lingual Transferability of Monolingual Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Nicholas Asher, Swarnadeep Bhar, Akshay Chaturvedi, Julie Hunter, and Soumya Paul. 2023. Limits for learning with language models. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 236–248, Toronto, Canada. Association for Computational Linguistics.

Temirlan Auyespek, Thomas Mach, and Assylbekov Zhenisbek. 2022. Hyperbolic Embedding for Finding Syntax in BERT. In *International Conference of the Italian Association for Artificial Intelligence, AIxIA 2021 DP - Virtual, Online*, volume 3078.

Geoff Bacon and Terry Regier. 2019. Does BERT agree? Evaluating knowledge of structure dependence through agreement relations. *arXiv preprint*. ArXiv:1908.09892.

Jiangang Bai, Yujing Wang, Yiren Chen, Yaming Yang, Jing Bai, Jing Yu, and Yunhai Tong. 2021. Syntax-BERT: Improving Pre-trained Transformers with Syntax Trees. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3011–3020, Online. Association for Computational Linguistics.

Xuefeng Bai, Jialong Wu, Yulong Chen, Zhongqing Wang, and Yue Zhang. 2023. Constituency Parsing using LLMs. *arXiv preprint*. ArXiv:2310.19462.

Irene Baucells, Javier Aula-Blasco, Iria de Dios-Flores, Silvia Paniagua Suárez, Naiara Perez, Anna Salles, Susana Sotelo Docio, Júlia Falcão, Jose Javier Saiz, Robiert Sepulveda Torres, Jeremy Barnes, Pablo Gamallo, Aitor Gonzalez-Agirre, German Rigau, and Marta Villegas. 2025. IberoBench: A benchmark for LLM evaluation in Iberian languages. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10491–10519, Abu Dhabi, UAE. Association for Computational Linguistics.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Leonard Bereska and Stratis Gavves. 2024. Mechanistic interpretability for AI safety - a review. *Transactions on Machine Learning Research*. Survey Certification, Expert Certification.

Johannes Bjerva and Isabelle Augenstein. 2021. Does Typological Blinding Impede Cross-Lingual Sharing? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 480–486, Online. Association for Computational Linguistics.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2023. Prompting Language Models for Linguistic Structure. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6649–6663, Toronto, Canada. Association for Computational Linguistics.

Gemma Boleda. 2025. Llms as a synthesis between symbolic and continuous approaches to language. *Preprint*, arXiv:2502.11856.

Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. 2025. Large Language Models

Share Representations of Latent Grammatical Concepts Across Typologically Diverse Languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6131–6150, Albuquerque, New Mexico. Association for Computational Linguistics.

Bastian Bunzeck and Sina Zarrieß. 2025. Subword models struggle with word learning, but surprisal hides it. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 286–300, Vienna, Austria. Association for Computational Linguistics.

Necva Bölücü and Burcu Can. 2022. Analysing Syntactic and Semantic Features in Pre-trained Language Models in a Fully Unsupervised Setting. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 19–31, New Delhi, India. Association for Computational Linguistics.

Steven Cao, Nikita Kitaev, and Dan Klein. 2019. Multilingual Alignment of Contextual Word Representations.

Hande Celikkanat, Sami Virpioja, Jörg Tiedemann, and Marianna Apidianaki. 2020. Controlling the Imprint of Passivization and Negation in Contextualized Representations. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 136–148, Online. Association for Computational Linguistics.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Ben Bergen. 2024. When Is Multilinguality a Curse? Language Modeling for 250 High- and Low-Resource Languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4074–4096, Miami, Florida, USA. Association for Computational Linguistics.

Tyler A. Chang and Benjamin K. Bergen. 2022. Word acquisition in neural language models. *Transactions of the Association for Computational Linguistics*, 10:1–16.

Tyler A. Chang and Benjamin K. Bergen. 2023. Language model behavior: A comprehensive survey. *Preprint*, arXiv:2303.11504.

Tyler A. Chang, Yifan Xu, Weijian Xu, and Zhuowen Tu. 2021. Convolutions and self-attention: Reinterpreting relative positions in pre-trained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4322–4333, Online. Association for Computational Linguistics.

Rui P. Chaves and Stephanie N. Richter. 2021. Look at that! BERT can be easily distracted from paying attention to morphosyntax. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 28–38, Online. Association for Computational Linguistics.

Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L. Leavitt, and Naomi Saphra. 2023. Sudden Drops in the Loss: Syntax Acquisition, Phase Transitions, and Simplicity Bias in MLMs.

Boli Chen, Yao Fu, Guangwei Xu, Pengjun Xie, Chuanqi Tan, Mosha Chen, and Liping Jing. 2021. Probing {bert} in hyperbolic spaces. In *International Conference on Learning Representations*.

Xuanda Chen, Timothy O'Donnell, and Siva Reddy. 2024. When does word order matter and when doesn't it? *arXiv preprint*. ArXiv:2402.18838.

Jiali Cheng and Hadi Amiri. 2025. Linguistic Blind Spots of Large Language Models. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 1–17, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Daniil Cherniavskii, Eduard Tulchinskii, Vladislav Mikhailov, Irina Proskurina, Laida Kushnareva, Ekaterina Artemova, Serguei Barannikov, Irina Piontkovskaya, Dmitri Piontkovski, and Evgeny Burnaev. 2022. Acceptability Judgements via Examining the Topology of Attention Maps. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 88–107, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding Universal Grammatical Relations in Multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.

Rochelle Choenni and Ekaterina Shutova. 2022. Investigating Language Relationships in Multilingual Sentence Encoders Through the Lens of Linguistic Typology. *Computational Linguistics*, 48(3):635–672.

Leshem Choshen, Guy Hacohen, Daphna Weinshall, and Omri Abend. 2022. The Grammar-Learning Trajectories of Neural Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8281–8297, Dublin, Ireland. Association for Computational Linguistics.

Grzegorz Chrupała and Afra Alishahi. 2019. Correlating Neural and Symbolic Representations of Language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2952–2962, Florence, Italy. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What Does BERT Look at? An Analysis of BERT's Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Thomas Hikaru Clark, Clara Meister, Tiago Pimentel, Michael Hahn, Ryan Cotterell, Richard Futrell, and Roger Levy. 2023. A Cross-Linguistic Pressure for Uniform Information Density in Word Order. *Transactions of the Association for Computational Linguistics*, 11:1048–1065.

Louis Clouatre, Prasanna Parthasarathi, Amal Zouaq, and Sarath Chandar. 2022a. Local Structure Matters Most in Most Languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 285–294, Online only. Association for Computational Linguistics.

Louis Clouatre, Prasanna Parthasarathi, Amal Zouaq, and Sarath Chandar. 2022b. Local Structure Matters Most: Perturbation Study in NLU. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3712–3731, Dublin, Ireland. Association for Computational Linguistics.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, F. Viégas, and M. Wattenberg. 2019. Visualizing and Measuring the Geometry of BERT.

Simone Conia and Roberto Navigli. 2022. Probing for Predicate Argument Structures in Pretrained Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4622–4632, Dublin, Ireland. Association for Computational Linguistics.

Ionut Constantinescu, Tiago Pimentel, Ryan Cotterell, and Alex Warstadt. 2025. Investigating Critical Period Effects in Language Acquisition through Neural Language Models. *Transactions of the Association for Computational Linguistics*, 13:96–120.

Fahim Dalvi, Hassan Sajjad, Nadir Durrani, and Yonatan Belinkov. 2020. Analyzing Redundancy in Pretrained Transformer Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4908–4926, Online. Association for Computational Linguistics.

Forrest Davis and Marten van Schijndel. 2020. Discourse structure interacts with reference but not syntax in neural language models. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 396–407, Online. Association for Computational Linguistics.

Iria de Dios-Flores, Juan Garcia Amboage, and Marcos Garcia. 2023. Dependency resolution at the syntax-semantics interface: psycholinguistic and computational insights on control dependencies. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 203–222, Toronto, Canada. Association for Computational Linguistics.

Iria de Dios-Flores and Garcia González. 2022. A computational psycholinguistic evaluation of the syntactic abilities of Galician BERT models at the interface of dependency resolution and training time. *Procesamiento del lenguaje natural*, 69:15–26.

Andrea Gregor de Varda and Marco Marelli. 2023. Data-driven Cross-lingual Syntax: An Agreement Study with Massively Multilingual Models. *Computational Linguistics*, 49(2):261–299.

Vittoria Dentella, Fritz Günther, and Evelina Leivada. 2023. Systematic testing of three Language Models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National Academy of Sciences of the United States of America*, 120(51):e2309583120.

Vittoria Dentella, Fritz Günther, and Evelina Leivada. 2025. Language in vivo vs. in silico: Size matters but Larger Language Models still do not comprehend language on a par with humans due to impenetrable semantic reference. *PLOS ONE*, 20(7):e0327794.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*, pages 4171–4186, Minneapolis, MN.

Richard Diehl Martinez, Zébulon Goriely, Andrew Caines, Paula Buttery, and Lisa Beinborn. 2024. Mitigating Frequency Bias and Anisotropy in Language Model Pre-Training with Syntactic Smoothing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5999–6011, Miami, Florida, USA. Association for Computational Linguistics.

Keyu Duan, Yiran Zhao, Zhili Feng, Jinjie Ni, Tianyu Pang, Qian Liu, Tianle Cai, Longxu Dou, Kenji Kawaguchi, Anirudh Goyal, J. Zico Kolter, and Michael Qizhe Shieh. 2025a. Unnatural Languages Are Not Bugs but Features for LLMs.

Xufeng Duan, Zhaoqian Yao, Yunhao Zhang, Shaonan Wang, and Zhenguang G. Cai. 2025b. How Syntax Specialization Emerges in Language Models. *arXiv preprint*. ArXiv:2505.19548 version: 1.

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online. Association for Computational Linguistics.

Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. 2020. Analyzing Individual Neurons in Pre-trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4865–4880, Online. Association for Computational Linguistics.

Tiwalayo Eisape, Vineet Gangireddy, Roger Levy, and Yoon Kim. 2022. Probing for Incremental Parse States in Autoregressive Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2801–2813, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yassir El Mesbahi, Atif Mahmud, Abbas Ghaddar, Mehdi Rezagholizadeh, Phillippe Langlais, and Prasanna Parthasarathi. 2023. On the utility of enhancing BERT syntactic bias with Token Reordering Pretraining. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 165–182, Singapore. Association for Computational Linguistics.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Mohamed Elgaar and Hadi Amiri. 2023. Ling-CL: Understanding NLP Models through Linguistic Curricula. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13526–13542, Singapore. Association for Computational Linguistics.

Allyson Ettinger. 2020. What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Mohsen Fayyaz, Ehsan Aghazadeh, Ali Modarressi, Hosein Mohebbi, and Mohammad Taher Pilehvar. 2021. Not All Models Localize Linguistic Knowledge in the Same Place: A Layer-wise Probing on BERToids' Representations. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 375–388, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Steven Y. Feng, Noah D. Goodman, and Michael C. Frank. 2024. Is child-directed speech effective training data for language models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22055–22071, Miami, Florida, USA. Association for Computational Linguistics.

Javier Ferrando and Marta R. Costa-jussà. 2024. On the Similarity of Circuits across Languages: a Case Study on the Subject-verb Agreement Task. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10115–10125, Miami, Florida, USA. Association for Computational Linguistics.

Javier Ferrando and Elena Voita. 2024. Information Flow Routes: Automatically Interpreting Language Models at Scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17432–17445, Miami, Florida, USA. Association for Computational Linguistics.

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. 2021. Causal analysis of syntactic agreement mechanisms in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1828–1843, Online. Association for Computational Linguistics.

Marcos Garcia and Alfredo Crespo-Otero. 2022. A Targeted Assessment of the Syntactic Abilities of Transformer Models for Galician-Portuguese. In *Computational Processing of the Portuguese Language*, pages 46–56, Cham. Springer International Publishing.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online. Association for Computational Linguistics.

Lucas Georges Gabriel Charpentier and David Samuel. 2023. Not all layers are equally as important: Every Layer Counts BERT. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 238–252, Singapore. Association for Computational Linguistics.

Luke Gessler and Nathan Schneider. 2023. Syntactic Inductive Bias in Transformer Language Models: Especially Helpful for Low-Resource Languages? In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 238–253, Singapore. Association for Computational Linguistics.

Michael Ginn. 2024. Tree Transformers are an Ineffective Model of Syntactic Constituency. *arXiv preprint*. ArXiv:2411.16993.

Goran Glavaš and Ivan Vulić. 2021. Is Supervised Syntactic Parsing Beneficial for Language Understanding Tasks? An Empirical Investigation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3090–3104, Online. Association for Computational Linguistics.

Yoav Goldberg. 2019. Assessing BERT's Syntactic Abilities. *ArXiv*.

Google Labs. 2023. Notebooklm. https://notebooklm.google. AI-powered research and note-taking tool by Google Gemini.

Raffaele Guarasci, Stefano Silvestri, Giuseppe De Pietro, Hamido Fujita, and Massimo Esposito. 2022. BERT syntactic transfer: A computational experiment on Italian, French and English languages. *Computer Speech & Language*, 71:101261.

Raffaele Guarasci, Stefano Silvestri, Giuseppe De Pietro, Hamido Fujita, and Massimo Esposito. 2023. Assessing BERT's ability to learn Italian syntax: a study on null-subject and agreement phenomena. *Journal of Ambient Intelligence and Humanized Computing*, 14(1):289–303.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of NAACL*, pages 1195–1205, New Orleans, LA.

Ashim Gupta, Giorgi Kvernadze, and Vivek Srikumar. 2021a. Bert & family eat word salad: Experiments with text understanding. In *AAAI Conference on Artificial Intelligence*.

Vikram Gupta, Haoyue Shi, Kevin Gimpel, and Mrinmaya Sachan. 2021b. Deep Clustering of Text Representations for Supervision-free Probing of Syntax. *arXiv preprint*. ArXiv:2010.12784.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. 2023. Finding Neurons in a Haystack: Case Studies with Sparse Probing. *Transactions on Machine Learning Research*.

John T. Hale and Miloš Stanojević. 2024. Do LLMs learn a true syntactic universal? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17106–17119, Miami, Florida, USA. Association for Computational Linguistics.

Sophie Hao and Tal Linzen. 2023. Verb Conjugation in Transformers Is Determined by Linear Encodings of Subject Number. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4531–4539, Singapore. Association for Computational Linguistics.

Linyang He, Peili Chen, Ercong Nie, Yuanning Li, and Jonathan R. Brennan. 2024. Decoding Probing: Revealing Internal Linguistic Structures in Neural Language Models Using Minimal Pairs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4488–4497, Torino, Italia. ELRA and ICCL.

Linyang He, Ercong Nie, Helmut Schmid, Hinrich Schuetze, Nima Mesgarani, and Jonathan Brennan. 2025a. Large language models as neurolinguistic subjects: Discrepancy between performance and competence. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19284–19302, Vienna, Austria. Association for Computational Linguistics.

Linyang He, Ercong Nie, Helmut Schmid, Hinrich Schütze, Nima Mesgarani, and Jonathan Brennan. 2025b. Large language models as neurolinguistic subjects: Discrepancy in performance and competence for form and meaning. *Preprint*, arXiv:2411.07533.

Evan Hernandez and Jacob Andreas. 2021. The Low-Dimensional Linear Geometry of Contextualized Word Representations. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 82–93, Online. Association for Computational Linguistics.

Jack Hessel and Alexandra Schofield. 2021. How effective is BERT without word ordering? Implications for language understanding and data privacy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 204–211, Online. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A Structural Probe for Finding Syntax in Word Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

John Hewitt, John Thickstun, Christopher Manning, and Percy Liang. 2023. Backpack language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9103–9125, Toronto, Canada. Association for Computational Linguistics.

Ester Hlavnova and Sebastian Ruder. 2023. Empowering Cross-lingual Behavioral Testing of NLP Models with Typological Features. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7181–7198, Toronto, Canada. Association for Computational Linguistics.

Liwen Hou and David Smith. 2023. Detecting Syntactic Change with Pre-trained Transformer Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3564–3574, Singapore. Association for Computational Linguistics.

Yifan Hou and Mrinmaya Sachan. 2021. Bird's Eye: Probing for Linguistic Graph Structures with a Simple Information-Theoretic Approach. In *Proceedings of the 59th Annual Meeting of the Association*

for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1844–1859, Online. Association for Computational Linguistics.

Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R. Bowman. 2019. Do Attention Heads in BERT Track Syntactic Dependencies? *NY Academy of Sciences NLP, Dialog, and Speech Workshop*. ArXiv:1911.12246.

Jennifer Hu, Sherry Yong Chen, and Roger Levy. 2020a. A closer look at the performance of neural language models on reflexive anaphor licensing. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 323–333, New York, New York. Association for Computational Linguistics.

Jennifer Hu and Michael Frank. 2024. Auxiliary task demands mask the capabilities of smaller language models.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020b. A Systematic Assessment of Syntactic Generalization in Neural Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.

Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, 121(36):e2400917121.

Michael Y. Hu, Jackson Petty, Chuan Shi, William Merrill, and Tal Linzen. 2025. Between Circuits and Chomsky: Pre-pretraining on Formal Languages Imparts Linguistic Biases. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9691–9709, Vienna, Austria. Association for Computational Linguistics.

Jing Huang, Atticus Geiger, Karel D'Oosterlinck, Zhengxuan Wu, and Christopher Potts. 2023. Rigorously Assessing Natural Language Explanations of Neurons. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 317–331, Singapore. Association for Computational Linguistics.

Kuan-Jung Huang, Suhas Arehalli, Mari Kugemoto, Christian Muxica, Grusha Prasad, Brian Dillon, and Tal Linzen. 2024. Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, 137:104510.

Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. BabyBERTa: Learning More Grammar With Small-Scale Child-Directed Language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online. Association for Computational Linguistics.

Yusuke Ide, Yuto Nishida, Justin Vasselli, Miyu Oba, Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. 2025. How to Make the Most of LLMs' Grammatical Knowledge for Acceptability Judgments. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7416–7432, Albuquerque, New Mexico. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What Does BERT Learn about the Structure of Language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Jae-young Jo and Sung-Hyon Myaeng. 2020. Roles and Utilization of Attention Heads in Transformer-based Neural Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3404–3417, Online. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Da Ju, Hagen Blix, and Adina Williams. 2025. Domain Regeneration: How well do LLMs match syntactic properties of text domains? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2367–2388, Vienna, Austria. Association for Computational Linguistics.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2025. Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs. *Preprint*, arXiv:2504.02768.

Jaap Jumelet and Willem Zuidema. 2023. Feature Interactions Reveal Linguistic Structure in Language Models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8697–8712, Toronto, Canada. Association for Computational Linguistics.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-Lingual Ability of Multilingual BERT: An Empirical Study.

Patrick Kahardipraja, Olena Vyshnevska, and Sharid Loáiciga. 2020. Exploring span representations in

neural coreference resolution. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 32–41, Online. Association for Computational Linguistics.

Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. 2024. Mission: Impossible Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14691–14714, Bangkok, Thailand. Association for Computational Linguistics.

Buddhika Kasthuriarachchy, Madhu Chetty, Adrian Shatte, and Darren Walls. 2021. From General Language Understanding to Noisy Text Comprehension. *Applied Sciences*, 11(27).

Mary Kennedy. 2025a. Evidence of Generative Syntax in LLMs. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 377–396, Vienna, Austria. Association for Computational Linguistics.

Mary Kathryn Kennedy. 2025b. Evidence of Hierarchically-Complex Syntactic Structure Within BERT's Word Representations. *Society for Computation in Linguistics*, 8(1).

Corentin Kervadec, Francesca Franzon, and Marco Baroni. 2023. Unnatural language processing: How do language models handle machine-generated prompts? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14377–14392, Singapore. Association for Computational Linguistics.

Jong Myoung Kim, Young-Jun Lee, Yong-Jin Han, Ho-Jin Choi, and Sangkeun Jung. 2024. Does Incomplete Syntax Influence Korean Language Model? Focusing on Word Order and Case Markers.

Najoung Kim, Jatin Khilnani, Alex Warstadt, and Abdelrahim Qaddoumi. 2023. Reconstruction Probing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8240–8255, Toronto, Canada. Association for Computational Linguistics.

Najoung Kim and Paul Smolensky. 2021. Testing for Grammatical Category Abstraction in Neural Language Models. In *Proceedings of the Society for Computation in Linguistics 2021*, pages 467–470, Online. Association for Computational Linguistics.

Taeuk Kim, Jihun Choi, Daniel Edmiston, and Sang goo Lee. 2020. Are pre-trained language models aware of phrases? simple but strong baselines for grammar induction. In *International Conference on Learning Representations*.

Hassane Kissane, Achim Schilling, and Patrick Krauss. 2025. Probing Internal Representations of Multi-Word Verbs in Large Language Models. In *Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025)*, pages 7–13, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

Josef Klafka and Allyson Ettinger. 2020. Spying on Your Neighbors: Fine-grained Probing of Contextual Embeddings for Information about Surrounding Words. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4801–4811, Online. Association for Computational Linguistics.

Aomi Koyama, Masato Mita, Su-Youn Yoon, Yasufumi Takama, and Mamoru Komachi. 2025. Targeted Syntactic Evaluation for Grammatical Error Correction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21108–21125, Vienna, Austria. Association for Computational Linguistics.

Katarzyna Krasnowska-Kieraś and Alina Wróblewska. 2019. Empirical linguistic study of sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5729–5739, Florence, Italy. Association for Computational Linguistics.

Daria Kryvosheieva and Roger Levy. 2025. Controlled Evaluation of Syntactic Knowledge in Multilingual Language Models. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 402–413, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Artur Kulmizev and Joakim Nivre. 2022. Schrödinger's tree – on syntax and neural language models. *Frontiers in Artificial Intelligence*, Volume 5 - 2022.

Artur Kulmizev, Vinit Ravishankar, Mostafa Abdou, and Joakim Nivre. 2020. Do Neural Language Models Show Preferences for Syntactic Formalisms? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4077–4091, Online. Association for Computational Linguistics.

Adhiguna Kuncoro, Lingpeng Kong, Daniel Fried, Dani Yogatama, Laura Rimell, Chris Dyer, and Phil Blunsom. 2020. Syntactic structure distillation pretraining for bidirectional encoders. *Transactions of the Association for Computational Linguistics*, 8:776–794.

Jenny Kunz and Marco Kuhlmann. 2020. Classifier Probes May Just Learn from Linear Context Features. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5136–5146, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jenny Kunz and Marco Kuhlmann. 2021. Test Harder than You Train: Probing with Extrapolation Splits. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 15–25, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jenny Kunz and Marco Kuhlmann. 2022. Where Does Linguistic Information Emerge in Neural Language Models? Measuring Gains and Contributions across Layers. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4664–4676, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. 2021. Lower Perplexity is Not Always Human-Like. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5203–5217, Online. Association for Computational Linguistics.

Tatsuki Kuribayashi, Ryo Ueda, Ryo Yoshida, Yohei Oseki, Ted Briscoe, and Timothy Baldwin. 2024. Emergent Word Order Universals from Cognitively-Motivated Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14522–14543, Bangkok, Thailand. Association for Computational Linguistics.

Emanuele La Malfa, Matthew Wicker, and Marta Kwiatkowska. 2022. Emergent Linguistic Structures in Neural Networks are Fragile.

Vedang Lad, Wes Gurnee, and Max Tegmark. 2024. The remarkable robustness of LLMs: Stages of inference? In *ICML 2024 Workshop on Mechanistic Interpretability*.

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of ICML*, pages 2879–2888, Stockholm, Sweden.

Yair Lakretz, Théo Desbordes, Dieuwke Hupkes, and Stanislas Dehaene. 2022. Can Transformers Process Recursive Nested Constructions, Like Humans? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3226–3232, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Andrew Lampinen. 2024. Can Language Models Handle Recursively Nested Grammatical Structures? A Case Study on Comparing Models and Humans. *Computational Linguistics*, 50(3):1441–1476.

Karim Lasri, Alessandro Lenci, and Thierry Poibeau. 2022a. Does BERT really agree ? Fine-grained Analysis of Lexical Dependence on a Syntactic Task. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2309–2315, Dublin, Ireland. Association for Computational Linguistics.

Karim Lasri, Alessandro Lenci, and Thierry Poibeau. 2022b. Word Order Matters When You Increase Masking. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1808–1815, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Karim Lasri, Olga Seminck, Alessandro Lenci, and Thierry Poibeau. 2022c. Subject Verb Agreement Error Patterns in Meaningless Sentences: Humans vs. BERT. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 37–43, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, 41(5):1202–1241.

Jonghyun Lee and Jeong-Ah Shin. 2023. Decoding BERT's Internal Processing of Garden-Path Structures through Attention Maps. *Korean Journal of English Language and Linguistics*, 23:461–481.

Jonghyun Lee, Jeong Ah Shin, and Myung Kwan Park. 2022. (AL)BERT Down the Garden Path: Psycholinguistic Experiments for Pre-trained Language Models. *Korean Journal of English Language and Linguistics*, 22:1033–1050.

Soo-Hwan Lee and Sebastian Schuster. 2022. Can language models capture syntactic associations without surface cues? A case study of reflexive anaphor licensing in English control constructions. In *Proceedings of the Society for Computation in Linguistics 2022*, pages 206–211, online. Association for Computational Linguistics.

Cara Su-Yi Leong and Tal Linzen. 2024. Testing learning hypotheses using neural networks by manipulating learning data. *Preprint*, arXiv:2407.04593.

Michael Lepori and R. Thomas McCoy. 2020. Picking BERT's Brain: Probing for Linguistic Dependencies in Contextualized Embeddings Using Representational Similarity Analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3637–3651, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Andrew Li, Xianle Feng, Siddhant Narang, Austin Peng, Tianle Cai, Raj Sanjay Shah, and Sashank Varma. 2024. Incremental Comprehension of Garden-Path Sentences by Large Language Models: Semantic Interpretation, Syntactic Re-Analysis, and Attention. *arXiv preprint*. ArXiv:2405.16042.

Bai Li, Zining Zhu, Guillaume Thomas, Frank Rudzicz, and Yang Xu. 2022a. Neural reality of argument structure constructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7410–7423, Dublin, Ireland. Association for Computational Linguistics.

Bai Li, Zining Zhu, Guillaume Thomas, Yang Xu, and Frank Rudzicz. 2021a. How is BERT surprised? Layerwise detection of linguistic anomalies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4215–4228, Online. Association for Computational Linguistics.

Bingzhi Li, Guillaume Wisniewski, and Benoit Crabbé. 2021b. Are Transformers a Modern Version of ELIZA? Observations on French Object Verb Agreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4599–4610, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiaoda Li, Ryan Cotterell, and Mrinmaya Sachan. 2022b. Probing via Prompting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1144–1157, Seattle, United States. Association for Computational Linguistics.

Tomasz Limisiewicz and David Mareček. 2020. Syntax representation in word embeddings and neural networks - a survey. *ArXiv*, abs/2010.01063.

Tomasz Limisiewicz and David Mareček. 2021. Introducing Orthogonal Constraint in Structural Probes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 428–442, Online. Association for Computational Linguistics.

Tomasz Limisiewicz, David Mareček, and Rudolf Rosa. 2020. Universal Dependencies According to BERT: Both More Specific and More General. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2710–2722, Online. Association for Computational Linguistics.

Boda Lin, Xinyi Zhou, Binghao Tang, Xiaocheng Gong, and Si Li. 2023. ChatGPT is a Potential Zero-Shot Dependency Parser. *arXiv preprint*. ArXiv:2310.16654.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open Sesame: Getting inside BERT's Linguistic Knowledge. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 241–253, Florence, Italy. Association for Computational Linguistics.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from Deep Learning. *Annual Review of Linguistics*, 7:195–212.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Zeyu Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A. Smith. 2021. Probing Across Time: What Does RoBERTa Know and When? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 820–842, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Charles Lovering, Rohan Jha, Tal Linzen, and Ellie Pavlick. 2021. Predicting inductive biases of pre-trained models. In *International Conference on Learning Representations*.

Ziyang Luo. 2021. Have Attention Heads in BERT Learned Constituency Grammar? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 8–15, Online. Association for Computational Linguistics.

Miguel López-Otal, Jorge Gracia, Jordi Bernad, Carlos Bobed, Lucía Pitarch-Ballesteros, and Emma Anglés-Herrero. 2025. Linguistic interpretability of transformer-based language models: a systematic review. *Preprint*, arXiv:2504.08001.

MA. Kelly, Y. Xu, J. Calvillo, and D. Reitter. 2020. Which Sentence Embeddings and Which LayersEncode Syntactic Structure? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 42.

Madhura Pande, Aakriti Budhraja, Preksha Nema, and Pratyush Kumar. 2021. The heads hypothesis: A unifying statistical approach towards understanding multi-headed attention in BERT. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35 of *15*, pages 13613–13621.

Kyle Mahowald. 2023. A Discerning Several Thousand Judgments: GPT-3 Rates the Article + Adjective + Numeral + Noun Construction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 265–273, Dubrovnik, Croatia. Association for Computational Linguistics.

Nikolay Malkin, Sameera Lanka, Pranav Goel, and Nebojsa Jojic. 2021. Studying word order through iterative shuffling. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10351–10366, Online and Punta

Cana, Dominican Republic. Association for Computational Linguistics.

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.

David Mareček and Rudolf Rosa. 2019. From Balustrades to Pierre Vinken: Looking for Syntax in Transformer Self-Attentions. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 263–275, Florence, Italy. Association for Computational Linguistics.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. 2024. Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models.

Samuel Marks and Max Tegmark. 2024. The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets.

Rowan Hall Maudslay and Ryan Cotterell. 2021. Do Syntactic Probes Probe Syntax? Experiments with Jabberwocky Probing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 124–131, Online. Association for Computational Linguistics.

Rowan Hall Maudslay, Josef Valvoda, Tiago Pimentel, Adina Williams, and Ryan Cotterell. 2020. A Tale of a Probe and a Parser. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7389–7395, Online. Association for Computational Linguistics.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

R. Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D. Hardy, and Thomas L. Griffiths. 2024. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121.

Thomas A. McGee and Idan A. Blank. 2024. Evidence Against Syntactic Encapsulation in Large Language Models. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46(0).

Alessio Miaschi, Chiara Alzetta, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2021. Probing tasks under pressure. In *Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-it 2021)*, pages 230–236, Milan, Italy. CEUR Workshop Proceedings.

Alessio Miaschi, Alzetta, Chiara, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2023. Testing the Effectiveness of the Diagnostic Probing Paradigm on Italian Treebanks. 3:144.

Alessio Miaschi, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2020a. Linguistic profiling of a neural language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 745–756, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alessio Miaschi and Felice Dell'Orletta. 2020. Contextual and Non-Contextual Word Embeddings: an in-depth Linguistic Investigation. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 110–119, Online. Association for Computational Linguistics.

Alessio Miaschi, Felice Dell'Orletta, and Giulia Venturi. 2024. Evaluating Large Language Models via Linguistic Profiling. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2835–2848, Miami, Florida, USA. Association for Computational Linguistics.

Alessio Miaschi, Gabriele Sarti, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. 2020b. Italian Transformers Under the Linguistic Lens. In Felice Dell'Orletta, Johanna Monti, and Fabio Tamburini, editors, *Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020 : Bologna, Italy, March 1-3, 2021*, Collana dell'Associazione Italiana di Linguistica Computazionale, pages 310–316. Accademia University Press, Torino.

Julian Michael, Jan A. Botha, and Ian Tenney. 2020. Asking without Telling: Exploring Latent Ontologies in Contextual Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6792–6812, Online. Association for Computational Linguistics.

James A. Michaelov, Catherine Arnett, Tyler A. Chang, and Benjamin K. Bergen. 2023. Structural Priming Demonstrates Abstract Grammatical Representations in Multilingual Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3703–3720, Singapore. Association for Computational Linguistics.

Vladislav Mikhailov, Oleg Serikov, and Ekaterina Artemova. 2021a. Morph Call: Probing Morphosyntactic Content of Multilingual Transformers. In *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, pages 97–121, Online. Association for Computational Linguistics.

Vladislav Mikhailov, Ekaterina Taktasheva, Elina Sigdel, and Ekaterina Artemova. 2021b. RuSentEval: Linguistic Source, Encoder Force! In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 43–65, Kiyv, Ukraine. Association for Computational Linguistics.

Raphaël Millière. 2024. Language models as models of language. *Preprint*, arXiv:2408.07144.

Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic Data Augmentation Increases Robustness to Inference Heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2339–2352, Online. Association for Computational Linguistics.

Kanishka Misra. 2022. minicons: Enabling Flexible Behavioral and Representational Analyses of Transformer Language Models. *arXiv preprint*. ArXiv:2203.13112.

Kanishka Misra and Kyle Mahowald. 2024. Language Models Learn Rare Phenomena from Less Rare Phenomena: The Case of the Missing AANNs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 913–929, Miami, Florida, USA. Association for Computational Linguistics.

Hosein Mohebbi, Ali Modarressi, and Mohammad Taher Pilehvar. 2021. Exploring the Role of BERT Token Representations to Explain Sentence Probing Results. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 792–806, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. Cross-Linguistic Syntactic Evaluation of Word Prediction Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online. Association for Computational Linguistics.

Aaron Mueller, Yu Xia, and Tal Linzen. 2022. Causal Analysis of Syntactic Agreement Neurons in Multilingual Language Models. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 95–109, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Shikhar Murty, Pratyusha Sharma, Jacob Andreas, and Christopher Manning. 2023. Grokking of Hierarchical Structure in Vanilla Transformers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 439–448, Toronto, Canada. Association for Computational Linguistics.

Alberto Muñoz-Ortiz, David Vilares, and Carlos Gómez-Rodríguez. 2023. Assessment of Pre-Trained Models Across Languages and Grammars. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 359–373, Nusa Dua, Bali. Association for Computational Linguistics.

Aleksandra Mysiak and Jacek Cyranka. 2023. Is German secretly a Slavic language? What BERT probing can tell us about language groups. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 86–93, Dubrovnik, Croatia. Association for Computational Linguistics.

Max Müller-Eberstein, Rob van der Goot, Barbara Plank, and Ivan Titov. 2023. Subspace Chronicles: How Linguistic Information Emerges, Shifts and Interacts during Language Model Training. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13190–13208, Singapore. Association for Computational Linguistics.

N. Atox and M. Clark. Evaluating Large Language Models through the Lens of Linguistic Proficiency and World Knowledge: A Comparative Study. *Authorea*, (August 27, 2024).

Ananjan Nandi, Christopher D Manning, and Shikhar Murty. 2025. Sneaking Syntax into Transformer Language Models with Tree Regularization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8006–8024, Albuquerque, New Mexico. Association for Computational Linguistics.

Anmol Nayak and Hari Prasad Timmapathini. 2021. Using Integrated Gradients and Constituency Parse Trees to explain Linguistic Acceptability learnt by BERT. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 80–85, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Benjamin Newman, Kai-Siang Ang, Julia Gong, and John Hewitt. 2021. Refining Targeted Syntactic Evaluation of Language Models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3710–3723, Online. Association for Computational Linguistics.

Dmitry Nikolaev and Sebastian Padó. 2023. The Universe of Utterances According to BERT. In *Proceedings of the 15th International Conference on Computational Semantics*, pages 99–105, Nancy, France. Association for Computational Linguistics.

Vassilina Nikoulina, Maxat Tezekbayev, Nuradil Kozhakhmet, Madina Babazhanova, Matthias Gallé,

and Zhenisbek Assylbekov. 2021. The Rediscovery Hypothesis: Language Models Need to Meet Linguistics. *Journal of Artificial Intelligence Research*, 72:1343–1384.

Jingcheng Niu, Wenjie Lu, Eric Corlett, and Gerald Penn. 2022a. Using Roark-Hollingshead Distance to Probe BERT's Syntactic Competence. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 325–334, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jingcheng Niu, Wenjie Lu, and Gerald Penn. 2022b. Does BERT Rediscover a Classical NLP Pipeline? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3143–3153, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Daisuke Oba, Naoki Yoshinaga, and Masashi Toyoda. 2021. Exploratory Model Analysis Using Data-Driven Neuron Representations. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 518–528, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Subba Reddy Oota, Manish Gupta, and Mariya Toneva. 2023. Joint processing of linguistic properties in brains and language models.

OpenAI. 2025. Chatgpt (gpt-4, gpt-5). `https://chat.openai.com`. Large language model.

Yulia Otmakhova, Karin Verspoor, and Jey Han Lau. 2022. Cross-linguistic Comparison of Linguistic Feature Encoding in BERT Models for Typologically Different Languages. In *Proceedings of the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 27–35, Seattle, Washington. Association for Computational Linguistics.

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, and 7 others. 2021. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372.

Isabel Papadimitriou, Ethan A. Chi, Richard Futrell, and Kyle Mahowald. 2021. Deep Subjecthood: Higher-Order Grammatical Features in Multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2522–2532, Online. Association for Computational Linguistics.

Isabel Papadimitriou, Richard Futrell, and Kyle Mahowald. 2022. When classifying arguments, BERT doesn't care about word order...except when it matters. In *Proceedings of the Society for Computation in Linguistics 2022*, pages 203–205, online. Association for Computational Linguistics.

Isabel Papadimitriou and Dan Jurafsky. 2023. Injecting structural hints: Using language models to study inductive biases in language learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8402–8413, Singapore. Association for Computational Linguistics.

Kwonsik Park, Myung Kwan Park, and Sanghoun Song. 2021. Deep learning can contrast the minimal pairs of syntactic data. *언어연구*, 38(2):395–424.

Abhinav Patil, Jaap Jumelet, Yu Ying Chiu, Andy Lapastora, Peter Shen, Lexie Wang, Clevis Willrich, and Shane Steinert-Threlkeld. 2024. Filtered Corpus Training (FiCT) Shows that Language Models Can Generalize from Indirect Evidence. *Transactions of the Association for Computational Linguistics*, 12:1597–1615.

Laura Pérez-Mayos, Alba Táboas García, Simon Mille, and Leo Wanner. 2021. Assessing the syntactic capabilities of transformer-based multilingual language models. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3799–3812, Online. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Erika Petersen and Christopher Potts. 2023. Lexical Semantics with Large Language Models: A Case Study of English "break". In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 490–511, Dubrovnik, Croatia. Association for Computational Linguistics.

Thang Pham, Trung Bui, Long Mai, and Anh Nguyen. 2021. Out of Order: How important is the sequential order of words in a sentence in Natural Language Understanding tasks? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1145–1160, Online. Association for Computational Linguistics.

Tiago Pimentel, Naomi Saphra, Adina Williams, and Ryan Cotterell. 2020a. Pareto Probing: Trading Off Accuracy for Complexity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3138–3153, Online. Association for Computational Linguistics.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell.

2020b. Information-Theoretic Probing for Linguistic Structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics.

Tiago Pimentel, Josef Valvoda, Niklas Stoehr, and Ryan Cotterell. 2022. The Architectural Bottleneck Principle. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11459–11472, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How Multilingual is Multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Jakob Prange, Nathan Schneider, and Lingpeng Kong. 2022. Linguistic Frameworks Go Toe-to-Toe at Neuro-Symbolic Language Modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4375–4391, Seattle, United States. Association for Computational Linguistics.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-Task Transfer Learning with Pretrained Language Models: When and Why Does It Work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online. Association for Computational Linguistics.

Giovanni Puccetti, Alessio Miaschi, and Felice Dell'Orletta. 2021. How Do BERT Embeddings Organize Linguistic Knowledge? In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 48–57, Online. Association for Computational Linguistics.

Laura Pérez-Mayos, Miguel Ballesteros, and Leo Wanner. 2021a. How much pretraining data do language models need to learn syntax? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1571–1582, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Laura Pérez-Mayos, Roberto Carlini, Miguel Ballesteros, and Leo Wanner. 2021b. On the evolution of syntactic information encoded by BERT's contextualized representations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2243–2258, Online. Association for Computational Linguistics.

Laura Pérez-Mayos, Alba Táboas García, Simon Mille, and Leo Wanner. 2021c. Assessing the Syntactic Capabilities of Transformer-based Multilingual Language Models. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3799–3812, Online. Association for Computational Linguistics.

Zhuang Qiu, Xufeng Duan, and Zhenguang G. Cai. 2025. Grammaticality representation in ChatGPT as compared to linguists and laypeople. *Humanities and Social Sciences Communications*, 12(1):1–15.

Shauli Ravfogel, Grusha Prasad, Tal Linzen, and Yoav Goldberg. 2021. Counterfactual Interventions Reveal the Causal Effect of Relative Clause Representations on Agreement Prediction. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 194–209, Online. Association for Computational Linguistics.

Vinit Ravishankar, Memduh Gökırmak, Lilja Øvrelid, and Erik Velldal. 2019. Multilingual Probing of Deep Pre-Trained Contextual Encoders. In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 37–47, Turku, Finland. Linköping University Electronic Press.

Vinit Ravishankar, Artur Kulmizev, Mostafa Abdou, Anders Søgaard, and Joakim Nivre. 2021. Attention Can Reflect Syntactic Structure (If You Let It). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3031–3045, Online. Association for Computational Linguistics.

Vinit Ravishankar and Joakim Nivre. 2022. The Effects of Corpus Choice and Morphosyntax on Multilingual Space Induction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4130–4139, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ryokan Ri and Yoshimasa Tsuruoka. 2022. Pretraining with Artificial Language: Studying Transferable Knowledge in Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7302–7315, Dublin, Ireland. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Rudolf Rosa and David Mareček. 2019. Inducing Syntactic Trees from BERT Representations. *arXiv preprint*. ArXiv:1906.11511.

Subhro Roy, Sam Thomson, Tongfei Chen, Richard Shin, Adam Pauls, Jason Eisner, and Benjamin Van Durme. 2023. BenchCLAMP: A Benchmark for Evaluating Language Models on Syntactic and Semantic Parsing.

Elena Sofia Ruzzetti, Federico Ranaldi, Felicia Logozzo, Michele Mastromattei, Leonardo Ranaldi, and Fabio Massimo Zanzotto. 2023. Exploring Linguistic Properties of Monolingual BERTs with Typological Classification among Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14447–14461, Singapore. Association for Computational Linguistics.

Soo Hyun Ryu and Richard Lewis. 2021. Accounting for Agreement Phenomena in Sentence Comprehension with Transformer Language Models: Effects of Similarity-based Interference on Surprisal and Attention. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 61–71, Online. Association for Computational Linguistics.

Samuel Rönnqvist, Jenna Kanerva, Tapio Salakoski, and Filip Ginter. 2019. Is Multilingual BERT Fluent in Language Generation? In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 29–36, Turku, Finland. Linköping University Electronic Press.

Devendra Sachan, Yuhao Zhang, Peng Qi, and William L. Hamilton. 2021. Do Syntax Trees Help Pre-trained Transformers Extract Information? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2647–2661, Online. Association for Computational Linguistics.

Hassan Sajjad, Nadir Durrani, Fahim Dalvi, Firoj Alam, Abdul Khan, and Jia Xu. 2022. Analyzing Encoded Concepts in Transformer Language Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3082–3101, Seattle, United States. Association for Computational Linguistics.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked Language Model Scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Felipe Salvatore, Marcelo Finger, and Roberto Hirata Jr. 2019. A logical-based corpus for cross-lingual evaluation. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 22–30, Hong Kong, China. Association for Computational Linguistics.

Laurent Sartran, Samuel Barrett, Adhiguna Kuncoro, Miloš Stanojević, Phil Blunsom, and Chris Dyer. 2022. Transformer Grammars: Augmenting Transformer Language Models with Syntactic Inductive Biases at Scale. *Transactions of the Association for Computational Linguistics*, 10:1423–1439.

Carolin M. Schuster and Simon Hegelich. 2022. From BERT's Point of View: Revealing the Prevailing Contextual Differences. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1120–1138, Dublin, Ireland. Association for Computational Linguistics.

R. Sevastjanova, A. Kalouli, C. Beck, H. Hauptmann, and M. El-Assady. 2022. LMFingerprints: Visual Explanations of Language Model Embedding Spaces through Layerwise Contextualization Scores. *Computer Graphics Forum*, 41(3):295–307.

Naomi Shapiro, Amandalynne Paullada, and Shane Steinert-Threlkeld. 2021. A multilabel approach to morphosyntactic probing. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4486–4524, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mandar Sharma, Nikhil Muralidhar, and Naren Ramakrishnan. 2023. Learning non-linguistic skills without sacrificing linguistic proficiency. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6178–6191, Toronto, Canada. Association for Computational Linguistics.

Judith Sieker and Sina Zarrieß. 2023. When Your Language Model Cannot Even Do Determiners Right: Probing for Anti-Presuppositions and the Maximize Presupposition! Principle. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 180–198, Singapore. Association for Computational Linguistics.

Pablo J. Diego Simon, Stéphane d'Ascoli, Emmanuel Chemla, Yair Lakretz, and Jean-Remi King. 2024. A polar coordinate system represents syntax in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Arabella Sinclair, Jaap Jumelet, Willem Zuidema, and Raquel Fernández. 2022. Structural Persistence in Language Models: Priming as a Window into Abstract Language Representations. *Transactions of the Association for Computational Linguistics*, 10:1031–1050.

Koustuv Sinha, Jon Gauthier, Aaron Mueller, Kanishka Misra, Keren Fuentes, Roger Levy, and Adina Williams. 2023. Language model acceptability judgements are not always robust to context. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6043–6063, Toronto, Canada. Association for Computational Linguistics.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021a. Masked Language Modeling and the Distributional Hypothesis: Order Word Matters Pre-training for Little. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2888–2913, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Koustuv Sinha, Amirhossein Kazemnejad, Siva Reddy, Joelle Pineau, Dieuwke Hupkes, and Adina Williams. 2022. The Curious Case of Absolute Position Embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4449–4472, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. 2021b. UnNatural Language Inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7329–7346, Online. Association for Computational Linguistics.

Taiga Someya and Yohei Oseki. 2023. JBLiMP: Japanese Benchmark of Linguistic Minimal Pairs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1581–1594, Dubrovnik, Croatia. Association for Computational Linguistics.

Taiga Someya, Ryo Yoshida, and Yohei Oseki. 2024. Targeted Syntactic Evaluation on the Chomsky Hierarchy. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15595–15605, Torino, Italia. ELRA and ICCL.

Taiga Someya, Ryo Yoshida, Hitomi Yanaka, and Yohei Oseki. 2025. Derivational Probing: Unveiling the Layer-wise Derivation of Syntactic Structures in Neural Language Models. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 93–104, Vienna, Austria. Association for Computational Linguistics.

Ionut-Teodor Sorodoc, Kristina Gulordava, and Gemma Boleda. 2020. Probing for Referential Information in Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4177–4189, Online. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 431 others. 2023. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Karolina Stanczak, Edoardo Ponti, Lucas Torroba Hennigen, Ryan Cotterell, and Isabelle Augenstein. 2022. Same Neurons, Different Languages: Probing Morphosyntax in Multilingual Pre-trained Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1589–1598, Seattle, United States. Association for Computational Linguistics.

Karolina Stańczak, Lucas Torroba Hennigen, Adina Williams, Ryan Cotterell, and Isabelle Augenstein. 2023. A Latent-Variable Model for Intrinsic Probing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13591–13599.

Dhanasekar Sundararaman, Vivek Subramanian, Guoyin Wang, Shijing Si, Dinghan Shen, Dong Wang, and Lawrence Carin. 2019. Syntax-Infused Transformer and BERT models for Machine Translation and Natural Language Understanding. *arXiv preprint*. ArXiv:1911.06156.

Vinitra Swamy, Angelika Romanou, and Martin Jaggi. 2021. Interpreting Language Models Through Knowledge Graph Extraction.

Anushka Swarup, Avanti Bhandarkar, Ronald Wilson, Tianyu Pan, and Damon Woodard. 2025. From syntax to semantics: Evaluating the impact of linguistic structures on LLM-based information extraction. In *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*, pages 36–48, Vienna, Austria. Association for Computational Linguistics.

Ekaterina Taktasheva, Vladislav Mikhailov, and Ekaterina Artemova. 2021. Shaking Syntactic Trees on the Sesame Street: Multilingual Probing with Controllable Perturbations. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 191–210, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. 2020. CxGBERT: BERT meets Construction Grammar. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4020–4032, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from

context? probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Tristan Thrush, Ethan Wilcox, and Roger Levy. 2020. Investigating Novel Verb Learning in BERT: Selectional Preference Classes and Alternation-Based Syntactic Generalization. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 265–275, Online. Association for Computational Linguistics.

Yuanhe Tian, Xu Li, Wei Wang, Guoqing Jin, Pengsen Cheng, and Yan Song. 2025. Large Language Models Enhanced by Plug and Play Syntactic Knowledge for Aspect-based Sentiment Analysis. *arXiv preprint*. ArXiv:2506.12991.

Hariprasad Timmapathini, Anmol Nayak, Sarathchandra Mandadi, Siva Sangada, Vaibhav Kesri, Karthikeyan Ponnalagu, and V. Venkoparao. 2021. Probing the SpanBERT Architecture to interpret Scientific Domain Adaptation Challenges for Coreference Resolution.

Lindia Tjuatja, Emmy Liu, Lori Levin, and Graham Neubig. 2023. Syntax and Semantics Meet in the "Middle": Probing the Syntax-Semantics Interface of LMs Through Agentivity. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 149–164, Toronto, Canada. Association for Computational Linguistics.

Lucas Torroba Hennigen, Adina Williams, and Ryan Cotterell. 2020. Intrinsic Probing through Dimension Selection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 197–216, Online. Association for Computational Linguistics.

Ke Tran, Arianna Bisazza, and Christof Monz. 2018. The importance of being recurrent for modeling hierarchical structure. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4731–4736, Brussels, Belgium. Association for Computational Linguistics.

Mycal Tucker, Tiwalayo Eisape, Peng Qian, Roger Levy, and Julie Shah. 2022. When Does Syntax Mediate Neural Language Model Performance? Evidence from Dropout Probes. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5393–5408, Seattle, United States. Association for Computational Linguistics.

Mycal Tucker, Peng Qian, and Roger Levy. 2021. What if This Modified That? Syntactic Interventions with Counterfactual Embeddings. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 862–875, Online. Association for Computational Linguistics.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering Language Models with Activation Engineering.

Marten van Schijndel, Aaron Mueller, and Tal Linzen. 2019. Quantity doesn't buy quality syntax with neural language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5831–5837, Hong Kong, China. Association for Computational Linguistics.

Elke Vandermeerschen and Miryam de Lhoneux. 2025. Supervised and Unsupervised Probing of Shortcut Learning: Case Study on the Emergence and Evolution of Syntactic Heuristics in BERT. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9592–9604, Vienna, Austria. Association for Computational Linguistics.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the Structure of Attention in a Transformer Language Model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy. Association for Computational Linguistics.

David Vilares, Michalina Strzyz, Anders Søgaard, and Carlos Gómez-Rodríguez. 2020. Parsing as Pretraining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9114–9121.

Andreas Waldis, Yotam Perlitz, Leshem Choshen, Yufang Hou, and Iryna Gurevych. 2024. Holmes \ensuremath\recorder A Benchmark to Assess the Linguistic Competence of Language Models. *Transactions of the Association for Computational Linguistics*, 12:1616–1647.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019a. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Daphne P. Wang, Mehrnoosh Sadrzadeh, Miloš Stanojević, Wing-Yee Chow, and Richard Breheny. 2025. Extracting structure from an LLM - how to improve on surprisal-based models of Human Language Processing. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4938–4944, Abu Dhabi, UAE. Association for Computational Linguistics.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 Small.

Yiwen Wang, Jennifer Hu, Roger Levy, and Peng Qian. 2021. Controlled evaluation of grammatical knowledge in Mandarin Chinese language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5604–5620, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019b. Cross-Lingual BERT Transformation for Zero-Shot Dependency Parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5721–5727, Hong Kong, China. Association for Computational Linguistics.

Alex Warstadt and Samuel R. Bowman. 2020. Can neural networks acquire a structural bias from raw linguistic data? *ArXiv*.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. 2019. Investigating BERT's Knowledge of Language: Five Analysis Methods with NPIs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2877–2887, Hong Kong, China. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020a. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R. Bowman. 2020b. Learning Which Features Matter: RoBERTa Acquires a Preference for Linguistic Generalizations (Eventually). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 217–235, Online. Association for Computational Linguistics.

Lucas Weber, Jaap Jumelet, Elia Bruni, and Dieuwke Hupkes. 2024. Interpretability of Language Models via Task Spaces. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4522–4538, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Dan Garrette, Tal Linzen, and Ellie Pavlick. 2021. Frequency effects on syntactic rule learning in transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 932–948, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Leonie Weissweiler, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze. 2022. The better your Syntax, the better your Semantics? Probing Pretrained Language Models for the English Comparative Correlative. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10859–10882, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Leonie Weissweiler, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze. 2023. Explaining pretrained language models' understanding of linguistic structures using construction grammar. *Frontiers in Artificial Intelligence*, 6:1225791.

Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. 2023. Should You Mask 15% in Masked Language Modeling? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2985–3000, Dubrovnik, Croatia. Association for Computational Linguistics.

Jennifer C. White and Ryan Cotterell. 2021. Examining the inductive bias of neural language models with artificial languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 454–463, Online. Association for Computational Linguistics.

Jennifer C. White, Tiago Pimentel, Naomi Saphra, and Ryan Cotterell. 2021. A Non-Linear Structural Probe. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 132–138, Online. Association for Computational Linguistics.

Sarah Wiegreffe and Naomi Saphra. 2024. Mechanistic? In *The 7th BlackboxNLP Workshop*.

Thilini Wijesiriwardene, Ruwan Wickramarachchi, Aishwarya Naresh Reganti, Vinija Jain, Aman Chadha, Amit Sheth, and Amitava Das. 2024. On the relationship between sentence analogy identification and sentence structure encoding in large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 451–457, St. Julian's, Malta. Association for Computational Linguistics.

Gijs Wijnholds. 2023. Assessing Monotonicity Reasoning in Dutch through Natural Language Inference. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1494–1500,

Dubrovnik, Croatia. Association for Computational Linguistics.

Ethan Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2022. *Learning Syntactic Structures from String Input*, pages 113–138.

Ethan Wilcox, Pranali Vani, and Roger Levy. 2021. A Targeted Assessment of Incremental Processing in Neural Language Models and Humans. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 939–952, Online. Association for Computational Linguistics.

Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. 2024. Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*, 55(4):805–848.

Ethan Gotlieb Wilcox, Michael Y. Hu, Aaron Mueller, Alex Warstadt, Leshem Choshen, Chengxu Zhuang, Adina Williams, Ryan Cotterell, and Tal Linzen. 2025. Bigger is not always better: The importance of human-scale language modeling for psycholinguistics. *Journal of Memory and Language*, 144:104650.

Shijie Wu and Mark Dredze. 2019. Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are All Languages Created Equal in Multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Zhaofeng Wu, Hao Peng, and Noah A. Smith. 2021. Infusing Finetuning with Semantic Dependencies. *Transactions of the Association for Computational Linguistics*, 9:226–242.

Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. 2020. Perturbed Masking: Parameter-free Probing for Analyzing and Interpreting BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4166–4176, Online. Association for Computational Linguistics.

Vicky Xefteri, Tim Vieira, Ryan Cotterell, and Afra Amini. 2025. Syntactic control of language models by posterior inference. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25350–25365, Vienna, Austria. Association for Computational Linguistics.

Ningyu Xu, Tao Gui, Ruotian Ma, Qi Zhang, Jingting Ye, Menghan Zhang, and Xuanjing Huang. 2022. Cross-Linguistic Syntactic Difference in Multilingual BERT: How Good is It and How Does It Affect Transfer? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8073–8092, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ningyu Xu, Qi Zhang, Jingting Ye, Menghan Zhang, and Xuanjing Huang. 2023. Are Structural Concepts Universal in Transformer Language Models? Towards Interpretable Cross-Lingual Generalization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13951–13976, Singapore. Association for Computational Linguistics.

Tianyang Xu, Tatsuki Kuribayashi, Yohei Oseki, Ryan Cotterell, and Alex Warstadt. 2025. Can language models learn typologically implausible languages?

Zenan Xu, Daya Guo, Duyu Tang, Qinliang Su, Linjun Shou, Ming Gong, Wanjun Zhong, Xiaojun Quan, Daxin Jiang, and Nan Duan. 2021. Syntax-Enhanced Pre-trained Model. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5412–5422, Online. Association for Computational Linguistics.

Takateru Yamakoshi, James McClelland, Adele Goldberg, and Robert Hawkins. 2023. Causal interventions expose implicit situation models for commonsense language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13265–13293, Toronto, Canada. Association for Computational Linguistics.

Hitomi Yanaka and Koji Mineshima. 2021. Assessing the Generalization Capacity of Pre-trained Language Models through Japanese Adversarial Natural Language Inference. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 337–349, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, and Kentaro Inui. 2020. Do neural models learn systematicity of monotonicity inference in natural language? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6105–6117, Online. Association for Computational Linguistics.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha Abzianidze, and Johan Bos. 2019. Can neural networks understand monotonicity reasoning? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 31–40, Florence, Italy. Association for Computational Linguistics.

Xiulin Yang, Tatsuya Aoyama, Yuekun Yao, and Ethan Wilcox. 2025. Anything Goes? A Crosslinguistic Study of (Im)possible Language Learning in LMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26058–26077, Vienna, Austria. Association for Computational Linguistics.

David Yi, James Bruno, Jiayu Han, Peter Zukerman, and Shane Steinert-Threlkeld. 2022. Probing for Understanding of English Verb Classes and Alternations in Large Pre-trained Language Models. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 142–152, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kayo Yin and Graham Neubig. 2022. Interpreting Language Models with Contrastive Explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 184–198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ryo Yoshida and Yohei Oseki. 2022. Composition, attention, or both? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5822–5834, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Charles Yu, Ryan Sie, Nicolas Tedeschi, and Leon Bergen. 2020. Word Frequency Does Not Predict Grammatical Knowledge in Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4040–4054, Online. Association for Computational Linguistics.

Zeyu Yun, Yubei Chen, Bruno Olshausen, and Yann LeCun. 2021. Transformer visualization via dictionary learning: contextualized embedding as a linear superposition of transformer factors. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 1–10, Online. Association for Computational Linguistics.

Jingyi Zhang, Gerard de Melo, Hongfei Xu, and Kehai Chen. 2023a. A Closer Look at Transformer Attention for Multilingual Translation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 496–506, Singapore. Association for Computational Linguistics.

Lining Zhang, Mengchen Wang, Liben Chen, and Wenxin Zhang. 2022. Probing GPT-3's Linguistic Knowledge on Semantic Tasks. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 297–304, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Meishan Zhang, Zhenghua Li, Guohong Fu, and Min Zhang. 2021a. Dependency-based syntax-aware word representations. *Artificial Intelligence*, 292:103427.

Yan Zhang, Zhaopeng Feng, Zhiyang Teng, Zuozhu Liu, and Haizhou Li. 2023b. How Well Do Text Embedding Models Understand Syntax? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9717–9728, Singapore. Association for Computational Linguistics.

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. 2021b. When Do You Need Billions of Words of Pretraining Data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1112–1125, Online. Association for Computational Linguistics.

Zhihao Zhang, Jun Zhao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Unveiling Linguistic Regions in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6228–6247, Bangkok, Thailand. Association for Computational Linguistics.

Haoyu Zhao, Abhishek Panigrahi, Rong Ge, and Sanjeev Arora. 2023. Do Transformers Parse while Predicting the Masked Word? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16513–16542, Singapore. Association for Computational Linguistics.

Yida Zhao, Hao Xve, Xiang Hu, and Kewei Tu. 2025. A Systematic Study of Compositional Syntactic Transformer Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7070–7083, Vienna, Austria. Association for Computational Linguistics.

Yiyun Zhao and Steven Bethard. 2020. How does BERT's attention change when you fine-tune? An analysis methodology and a case study in negation scope. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4729–4747, Online. Association for Computational Linguistics.

Jianyu Zheng and Ying Liu. 2022. Probing language identity encoded in pre-trained multilingual models: a typological view. *PeerJ Computer Science*, 8:e899.

Jianyu Zheng and Ying Liu. 2023. What does Chinese BERT learn about syntactic knowledge? *PeerJ. Computer Science*, 9:e1478.

Houquan Zhou, Yang Hou, Zhenghua Li, Xuebin Wang, Zhefeng Wang, Xinyu Duan, and Min Zhang. 2023. How Well Do Large Language Models Understand Syntax? An Evaluation by Asking Natural Language Questions. *arXiv preprint*. ArXiv:2311.08287.

Junru Zhou, Zhuosheng Zhang, Hai Zhao, and Shuailiang Zhang. 2020. LIMIT-BERT : Linguistics Informed Multi-Task BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4450–4461, Online. Association for Computational Linguistics.

Xinyu Zhou, Delong Chen, Samuel Cahyawijaya, Xufeng Duan, and Zhenguang Cai. 2025. Linguistic Minimal Pairs Elicit Linguistic Similarity in Large Language Models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6866–6888, Abu Dhabi, UAE. Association for Computational Linguistics.

Zining Zhu and Frank Rudzicz. 2020. An information theoretic view on selecting linguistic probes. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9251–9262, Online. Association for Computational Linguistics.

Julia Witte Zimmerman, Denis Hudon, Kathryn Cramer, Alejandro J. Ruiz, Calla Beauregard, Ashley Fehr, Mikaela Irene Fudolig, Bradford Demarest, Yoshi Meke Bird, Milo Z. Trujillo, Christopher M. Danforth, and Peter Sheridan Dodds. 2025. Tokens, the oft-overlooked appetizer: Large language models, the distributional hypothesis, and meaning. *arXiv preprint*. ArXiv:2412.10924.

A References

Table 1 lists the studies included and analyzed in our database. The entries are organized first by publication year and then sorted alphabetically by the authors' names.

Table 1: The 337 studies included in the database, sorted by year and alphabetically.

2018: Tran et al. (2018), Peters et al. (2018)
2019: Abdou et al. (2019), Bacon and Regier (2019), Cao et al. (2019), Chrupała and Alishahi (2019), Clark et al. (2019), Coenen et al. (2019), Goldberg (2019), Hewitt and Manning (2019), Htut et al. (2019), Jawahar et al. (2019), Krasnowska-Kieraś and Wróblewska (2019), Lin et al. (2019), Liu et al. (2019), Mareček and Rosa (2019), McCoy et al. (2019), Pires et al. (2019), Ravishankar et al. (2019), Rönnqvist et al. (2019), Rosa and Mareček (2019), Salvatore et al. (2019), Sundararaman et al. (2019), Tenney et al. (2019a), Tenney et al. (2019b), van Schijndel et al. (2019), Vig and Belinkov (2019), Wang et al. (2019b), Wang et al. (2019a), Warstadt et al. (2019), Wu and Dredze (2019), Yanaka et al. (2019)
2020: Alt et al. (2020), Artetxe et al. (2020), Celikkanat et al. (2020), Chi et al. (2020), Dalvi et al. (2020), Davis and van Schijndel (2020), Durrani et al. (2020), Dufter and Schütze (2020), Ettinger (2020), Torroba Hennigen et al. (2020), Hu et al. (2020a), Hu et al. (2020b), Jo and Myaeng (2020), K et al. (2020), Kahardipraja et al. (2020), MA. Kelly et al. (2020), Kim et al. (2020), Klafka and Ettinger (2020), Kulmizev et al. (2020), Kunz and Kuhlmann (2020), Lepori and McCoy (2020), Limisiewicz et al. (2020), Tayyar Madabushi et al. (2020), Manning et al. (2020), Maudslay et al. (2020), Miaschi and Dell'Orletta (2020), Miaschi et al. (2020b), Miaschi et al. (2020a), Michael et al. (2020), Min et al. (2020), Mueller et al. (2020), Pimentel et al. (2020b), Pimentel et al. (2020a), Pruksachatkun et al. (2020), Ribeiro et al. (2020), Salazar et al. (2020), Sorodoc et al. (2020), Thrush et al. (2020), Vilares et al. (2020), Warstadt et al. (2020a), Warstadt and Bowman (2020), Warstadt et al. (2020b), Wu and Dredze (2020), Wu et al. (2020), Yanaka et al. (2020), Yu et al. (2020), Zhao and Bethard (2020), Zhou et al. (2020), Zhu and Rudzicz (2020)
2021: Alleman et al. (2021), Bai et al. (2021), Bjerva and Augenstein (2021), Chang et al. (2021), Chen et al. (2021), Chaves and Richter (2021), Elazar et al. (2021), Finlayson et al. (2021) Fayyaz et al. (2021), Glavaš and Vulić (2021), Gupta et al. (2021a), Gupta et al. (2021b), Hernandez and Andreas (2021), Hessel and Schofield (2021), Hou and Sachan (2021), Huebner et al. (2021), Kasthuriarachchy et al. (2021), Kim and Smolensky (2021), Kuncoro et al. (2020), Kunz and Kuhlmann (2021), Kuribayashi et al. (2021), Li et al. (2021b), Li et al. (2021a), Limisiewicz and Mareček (2021), Liu et al. (2021), Lovering et al. (2021), Luo (2021), Madhura Pande et al. (2021), Malkin et al. (2021), Maudslay and Cotterell (2021), Miaschi et al. (2021), Mikhailov et al. (2021a), Mikhailov et al. (2021b), Mohebbi et al. (2021), Nayak and Timmapathini (2021), Newman et al. (2021), Nikoulina et al. (2021), Oba et al. (2021), Papadimitriou et al. (2021), Park et al. (2021), Pérez-Mayos et al. (2021c), Pérez-Mayos et al. (2021b), Pérez-Mayos et al. (2021a), Pham et al. (2021), Puccetti et al. (2021), Ravfogel et al. (2021), Ravishankar et al. (2021), Ryu and Lewis (2021), Sachan et al. (2021), Shapiro et al. (2021), Sinha et al. (2021a), Sinha et al. (2021b), Swamy et al. (2021), Taktasheva et al. (2021), Timmapathini et al. (2021), Tucker et al. (2021), Wei et al. (2021), Wang et al. (2021), White et al. (2021), White and Cotterell (2021), Wilcox et al. (2021), Wu et al. (2021), Xu et al. (2021), Yanaka and Mineshima (2021), Yun et al. (2021), Zhang et al. (2021a), Zhang et al. (2021b)

Continued on next page

Table 1: The 337 studies included in the database, sorted by year and alphabetically. (Continued)

2022: Abdou et al. (2022), Alajrami and Aletras (2022), Alzetta et al. (2022), Arps et al. (2022), Aoyama and Schneider (2022) Auyespek et al. (2022), Bölücü and Can (2022), Chang and Bergen (2022), Cherniavskii et al. (2022), Choenni and Shutova (2022), Choshen et al. (2022), Clouatre et al. (2022a), Clouatre et al. (2022b), Conia and Navigli (2022), de Dios-Flores and González (2022), Eisape et al. (2022), Garcia and Crespo-Otero (2022), Guarasci et al. (2022), Kunz and Kuhlmann (2022), La Malfa et al. (2022), Lakretz et al. (2022), Lasri et al. (2022a), Lasri et al. (2022c), Lasri et al. (2022b), Lee et al. (2022), Lee and Schuster (2022), Li et al. (2022a), Li et al. (2022b), Misra (2022), Mueller et al. (2022), Niu et al. (2022b), Niu et al. (2022a), Otmakhova et al. (2022), Papadimitriou et al. (2022), Pimentel et al. (2022), Prange et al. (2022), Ravishankar and Nivre (2022), Ri and Tsuruoka (2022), Sajjad et al. (2022), Sartran et al. (2022), Schuster and Hegelich (2022), Sevastjanova et al. (2022), Sinclair et al. (2022), Sinha et al. (2022), Stanczak et al. (2022), Tucker et al. (2022), Wang et al. (2022), Weissweiler et al. (2022), Wilcox et al. (2022), Xu et al. (2022), Yi et al. (2022), Yin and Neubig (2022), Yoshida and Oseki (2022), Zhang et al. (2022), Zheng and Liu (2022)
2023: Amini et al. (2023), Asher et al. (2023), Bai et al. (2023), Blevins et al. (2023), Chen et al. (2023), Georges Gabriel Charpentier and Samuel (2023), Clark et al. (2023), de Dios-Flores et al. (2023), de Varda and Marelli (2023), Dentella et al. (2023), El Mesbahi et al. (2023), Elgaar and Amiri (2023), Gessler and Schneider (2023), Guarasci et al. (2023), Gurnee et al. (2023), Hao and Linzen (2023), Hlavnova and Ruder (2023), Hewitt et al. (2023), Hou and Smith (2023), Hu and Levy (2023), Huang et al. (2023), Jumelet and Zuidema (2023), Kervadec et al. (2023), Kim et al. (2023), Lee and Shin (2023), Lin et al. (2023), Mahowald (2023), Miaschi et al. (2023), Michaelov et al. (2023), Müller-Eberstein et al. (2023), Muñoz-Ortiz et al. (2023), Murty et al. (2023), Mysiak and Cyranka (2023), Nikolaev and Padó (2023), Oota et al. (2023), Papadimitriou and Jurafsky (2023), Petersen and Potts (2023), Roy et al. (2023), Ruzzetti et al. (2023), Sharma et al. (2023), Sieker and Zarrieß (2023), Sinha et al. (2023), Someya and Oseki (2023), Srivastava et al. (2023), Stańczak et al. (2023), Tjuatja et al. (2023), Weissweiler et al. (2023), Wettig et al. (2023), Wijnholds (2023), Xu et al. (2023), Yamakoshi et al. (2023), Zhang et al. (2023a), Zhang et al. (2023b), Zhao et al. (2023), Zheng and Liu (2023), Zhou et al. (2023)
2024: Acs et al. (2024), Allen-Zhu and Li (2024), Ambridge and Blything (2024), Arehalli and Linzen (2024), Arora et al. (2024), Arps et al. (2024), N. Atox and M. Clark, Chang et al. (2024), Chen et al. (2024), Diehl Martinez et al. (2024), Simon et al. (2024), Feng et al. (2024), Ferrando and Voita (2024), Ferrando and Costa-jussà (2024), Ginn (2024), Hale and Stanojević (2024), He et al. (2024), Hu and Frank (2024), Hu et al. (2024), Huang et al. (2024), Kallini et al. (2024), Kim et al. (2024), Kuribayashi et al. (2024), Lampinen (2024), Leong and Linzen (2024), Li et al. (2024), Marks and Tegmark (2024), Marks et al. (2024), McCoy et al. (2024), McGee and Blank (2024), Miaschi et al. (2024), Misra and Mahowald (2024), Patil et al. (2024), Someya et al. (2024), Turner et al. (2024), Waldis et al. (2024), Weber et al. (2024), Wijesiriwardene et al. (2024), Wilcox et al. (2024), Zhang et al. (2024)
2025: Acs et al. (2024) Agarwal et al. (2025), AlKhamissi et al. (2025), Alzetta et al. (2025), Ahuja et al. (2025), Aljaafari et al. (2025), Aoyama and Wilcox (2025), Brinkmann et al. (2025), Bunzeck and Zarrieß (2025), Cheng and Amiri (2025), Constantinescu et al. (2025), Dentella et al. (2025), Duan et al. (2025b), Duan et al. (2025a), He et al. (2025a), Hu et al. (2025), Ide et al. (2025), Ju et al. (2025), Jumelet et al. (2025), Kennedy (2025a), Kennedy (2025b), Kissane et al. (2025), Koyama et al. (2025), Kryvosheieva and Levy (2025), Nandi et al. (2025), Qiu et al. (2025), Someya et al. (2025), Swarup et al. (2025), Tian et al. (2025), Vandermeerschen and de Lhoneux (2025), Wang et al. (2025), Wilcox et al. (2025), Xefteri et al. (2025), Xu et al. (2025), Yang et al. (2025), Zhao et al. (2025), Zhou et al. (2025), Zimmerman et al. (2025)

B Annotation dimensions

This section gives more details on the annotation dimensions used in the database and lists the specific COLUMN NAME. Additional information on variable levels and coding schemes will be made publicly available in the repository's README.

B.1 Meta-Data

We publish the main information about the study (ID, SHORT_REF, URL, YEAR, TITLE, ABSTRACT), and information of its review process:

Source for the article: REVIEW, indicates where the article was sourced from with chronological line of adding papers. Tracks the order of inclusion from different sources.

Exclusion grounds: EXCLUSION_GROUNDS, indicates the reason for exclusion.

B.2 Model-related information

Model type: MODEL_TYPE, categorized as:

- **Bidirectional:** At least one bidirectional model is featured (e.g., BERT, LSTMs).
- **Causal:** At least one causal model is featured (e.g., GPT, LSTMs).
- **Both:** Both model types are featured (e.g., comparative experiments with BERT, GPT, and LSTM).
- **Other:** Cases with neither causal nor bidirectional models, either excluded or meeting special inclusion criteria, e.g., specific analysis of encoder component of a Transformer architecture.

Model name and post-processing: MODEL_NAME_COMMENT, for the comment string, as specific as possible it is stated in the paper. MODEL_NAME, we post-process this string into a list of dictionaries, regard the following example:

1. Name as reported in the paper, e.g., *bert_base_uncased*.
2. Name enriched with post processing script: e.g., *family* ="BERT" plus mention of size, BERT-base.
3. Final example: list({*string* ="bert_base_uncased", *family* ="BERT", *size* = base, *case* = uncased}).

Fine-tuned: FINE_TUNED, indicates whether fine-tuning was performed.

Language coverage: MODEL_LANGUAGE_COVERAGE

- **Monolingual:** One language from scratch or continually pretrained (e.g., GPT2, BERT, ruBERT).
- **Multilingual:** Trained on multiple languages simultaneously from scratch (e.g., mBERT, GPT3).
- **Both:** Studies comparing monolingual and multilingual models.

B.3 Experimental design

Syntactic phenomena:

- **General:** SYNTACTIC_PHENOMENA, bottom-up annotation, starting from general to more specific cases. See Table 2.
- **Specific:** SPECIFIC_PHENOMENA, as referred to by authors.

Languages tested and post-processing: LANGUAGES_TESTED_COMMENT, for the comment string, as specific as possible it is stated in the paper. LANGUAGES_TESTED, we post-process this string into a list of all languages used in the experiments. LANGUAGES_TESTED_ENG_INCLUDED, indicates whether English was tested. LANGUAGES_TESTED_MONOLINGUAL_ENG, indicates whether only English was tested. LANGUAGES_TESTED_LANG_COUNT, number of tested languages.

Fine-tuning materials: FINE_TUNING_MATERIALS, resources used for fine-tuning in the experiments.

Experimental materials: EXPERIMENTAL_MATERIALS_COMMENT, materials as referenced by authors and summarized through AI-based workflow. EXPERIMENTAL_MATERIALS, annotated in post-processing to indicate whether they are naturalistic language or controlled stimuli, both, or "unknown" (if under-specified/referring to other's paper dataset).

B.4 Interpretability techniques

General: INTERP_CATEGORY, categorized as:

- **Behavioral:** Methods based on last layer input–output behavior.

Phenomena	Description
Benchmarks	Aggregated evaluation suites combining multiple syntactic phenomena (e.g. BLiMP, CoLA, SyntaxGym), often reporting composite performance metrics rather than phenomenon-specific analyses.
Dependency parsing & constituency	Structural representations of sentence form, including dependency relations, constituency trees, and syntactic functions (e.g., subject vs. object), without targeting hierarchical depth or processing complexity.
Hierarchical structure	Properties of abstract hierarchical organization independent of specific formalisms, including tree depth, recursion, center embedding, attachment ambiguity, and structural complexity.
Word order	Constraints on the linear ordering of words and phrases, including canonical and non-canonical orders, scrambling, and variation in surface order.
Agreement & feature-checking	Morphosyntactic dependencies involving grammatical features such as number, gender, person, and tense, typically requiring local or semi-local syntactic constraints.
POS & other lexical properties	Lexical-level categories and features, such as part of speech, verbal subcategorization, or lexical aspect; excludes tasks like named entity recognition that are primarily semantic or discourse-driven.
Negation & NPIs	Licensing and interpretation of negation and negative polarity items, involving interactions between syntax, semantics, and scope.
Binding & coreference	Constraints on anaphora and antecedent resolution, including reflexives, pronouns, and coreference relations, typically requiring hierarchical and discourse-sensitive representations.
Ellipsis & filler–gap dependencies	Long-distance dependencies such as wh-movement, gap filling, and ellipsis resolution, requiring sensitivity to abstract structure beyond linear proximity.
Typology	Cross-linguistic variation in syntactic patterns and constraints, including agreement systems, and constructional differences across languages.
Argument structure & constructions	Systematic syntactic patterns linking form and meaning, including argument realization, alternations (e.g. active/passive, dative alternation), relative clauses, and other constructional schemas at the syntax–semantics interface.

Table 2: The 11 Linguistic phenomena of our survey with brief definitions, which we annotated and organized from general Benchmark, to those focusing on syntactic structure and surface-level form (blue-purple tones) to form-and-function-related categories.

- **Probing:** Methods training a classifier on top of a given TLM layer.
- **Mechanistic:** Methods investigating internal dynamics, may incorporate behavioral or probing elements.

Specific: INTERP_METHOD, specific techniques description, named as in the original paper.

Probing: PROBING, classification of results to claim evidence for syntactic representations in the TLM.

Location: PROBING_LOCATION, specific location article states where performance peaks within the model.

Mechanistic Interpretability: We use the taxonomy introduced by Bereska and Gavves (2024) to characterize key dimensions: We annotated each mechanistic paper, 140 in total, along these dimensions:

- **Method:** MECHANISTIC_METHOD, string as general as possible as it is stated in the paper.
- **Causal Nature:** MECHANISTIC_CAUSAL NATURE, ranges from purely observational to interventional approaches.
- **Phase:** MECHANISTIC_PHASE, the phase distinguishes post-hoc from intrinsic approaches applied during model development.
- **Locality:** MECHANISTIC_LOCALITY, describes the level of analysis, ranging from individual components (e.g., neurons, attention heads) to full model architectures (e.g., residual stream).

Syntactic Awareness: MECHANISTIC_SYNTACTIC EVIDENCE/ AWARENESS captures whether mechanistic interpretability studies claim evidence of syntactic awareness of TLMs.

B.5 Findings

Main findings on syntactic knowledge: MAIN_FINDINGS, summarizes the findings of studies on how TLMs handle syntax, and their syntactic knowledge.

C AI-Based Workflow for Information Extraction

During the review process, we developed a semi-automated strategy combining NotebookLM (Google Labs, 2023) and ChatGPT (OpenAI,

Category	Question
Model information	
- Model type	*Which model do they test and which languages, do they do finetuning?*
Experimental design	
- Syntactic phenomena	*What are the "Syntactic phenomena, the Specific: list set as referred to by authors"*
- Experimental materials	*What is the "Experimental materials: naturalistic language, controlled stimuli"*
Interpretability techniques	
- Method name	*What is the Specific method name as referred to by the authors?*
Findings	
- Main findings on syntactic knowledge	*What are the main findings, especially on syntactic knowledge in the LM?*

Table 3: Questions used for AI-based information extraction.

2025) for structured information extraction on 5 variables, model type, syntactic phenomena, experimental materials, method name, and main syntactic findings, listed in Table 3.

Each paper was first analysed using NotebookLM (Google Labs, 2023) by posing a series of targeted questions to extract relevant details for these categories. The responses generated by NotebookLM were then summarized into a coherent paragraph using ChatGPT (OpenAI, 2025),[10] which we subsequently reviewed and used to populate the database. We specifically chose to prompt for paragraph-style summaries to maintain compact, integrated information for the database, rather than fragmented notes.

See Fig. 22 for an example. The workflow was iteratively refined during the course of the review process until it stabilized; starting at paper ID 120, this approach was applied without further adaptation. Following the validation of our approach, as described in the next section, we decided to publish the "main_findings" column of the database exclusively with LLM-generated output.

C.1 Validation of the Approach

To assess the reliability of our semi-automated information extraction strategy, we compared the AI-assisted summaries against the first, manually collected information ("gold" set, $n = 91$). Our evaluation confirmed that the semi-automated workflow captured the exact key details as the manual article review consistently, while improving efficiency. See Fig. 23 for an example. The summaries were AI-rated by ChatGPT (OpenAI, 2025) on a scale from 0 (completely unrelated) to 5 (equivalent), with a median score of $\bar{x} = 4.1$. The AI-assisted summaries tend to be more detailed and insightful for keyword searches, providing richer context. By contrast, the manually collected information was generally shorter, more general, and abstract, often relying on direct quotations and omitting specific details such as model comparison, task descriptions, and numeric results.

C.2 Limitations of Workflow

Some linguistic phenomena referenced in articles pertained to related work rather than the studies' own experiments, which can introduce ambiguity in AI-assisted summaries. During the manual annotation of the general syntactic phenomena, we identified and corrected a few such cases, but minor residual errors may remain.

C.3 Illustrative Figures

We provide screenshots from the workflow and interactions with the models in Fig. 22 and Fig. 23.

D Additional Results to Section 3

D.1 Additional Results to the Overview

We plot the coverage of model type and TLMs language coverage in Fig. 8. Most work focuses on a single model paradigm, primarily bidirectional models (59%) or causal models (28%), with only a small fraction directly comparing both (13%); similarly, evaluations largely target monolingual

[10] August 2025 release, incorporating GPT-4 and GPT-5 models

Phenomenon	N	Acceptable Example	Unacceptable Example
ANAPHOR AGR.	2	*Many girls insulted themselves.*	*Many girls insulted herself.*
ARG. STRUCTURE	9	*Rose wasn't disturbing Mark.*	*Rose wasn't boasting Mark.*
BINDING	7	*Carlos said that Lori helped him.*	*Carlos said that Lori helped himself.*
CONTROL/RAISING	5	*There was bound to be a fish escaping.*	*There was unable to be a fish escaping.*
DET.-NOUN AGR.	8	*Rachelle had bought that chair.*	*Rachelle had bought that chairs.*
ELLIPSIS	2	*Anne's doctor cleans one important book and Stacey cleans a few.*	*Anne's doctor cleans one book and Stacey cleans a few important.*
FILLER-GAP	7	*Brett knew what many waiters find.*	*Brett knew that many waiters find.*
IRREGULAR FORMS	2	*Aaron broke the unicycle.*	*Aaron broken the unicycle.*
ISLAND EFFECTS	8	*Whose hat should Tonya wear?*	*Whose should Tonya wear hat?*
NPI LICENSING	7	*The truck has clearly tipped over.*	*The truck has ever tipped over.*
QUANTIFIERS	4	*No boy knew fewer than six guys.*	*No boy knew at most six guys.*
SUBJECT-VERB AGR.	6	*These casseroles disgust Kayla.*	*These casseroles disgusts Kayla.*

Figure 7: Table from Warstadt et al. (2020a); minimal pairs from each of the 12 linguistic phenomenon categories covered by BLiMP. Column *N* records the number of different subphenomena exemplifying the relevant general phenomenon, 67 in total.

TLMs (66%), while multilingual (20%) and both monolingual–multilingual settings (14%) remain underexplored.

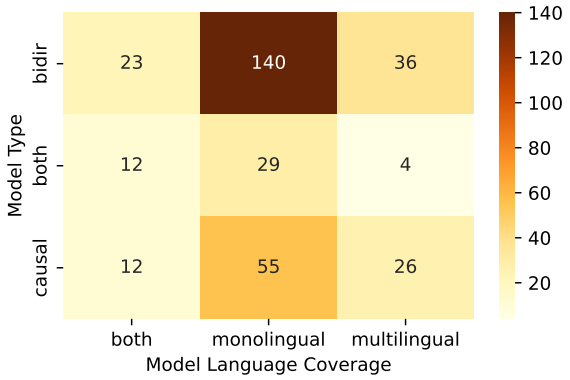


Figure 8: Count distribution by model type and language coverage.

We identified 54 unique TLMs and 609 non-duplicate model instances in the database; models were counted once per study and grouped by family or model name when appropriate (e.g., studies comparing BERT-base, BERT-large, or cased/uncased variants contributed a single BERT count). Fig. 9 shows that most of the results in our database focus on a single model, BERT, which accounts for 30% of all TLMS results. BERT and its variants together represent 58% of the datapoints, while the four most extensively studied models account for 62% overall.

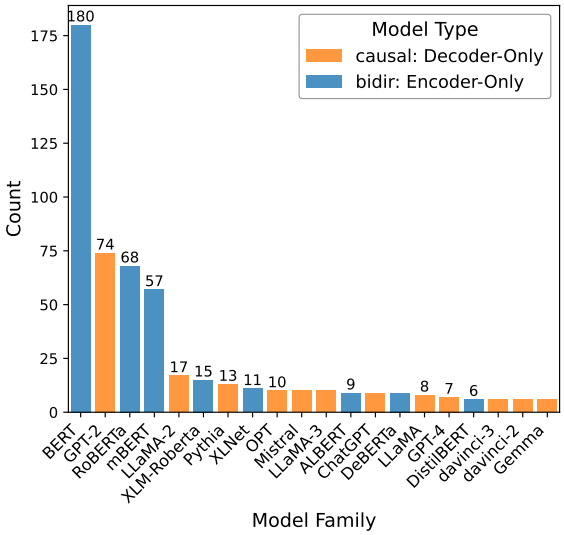


Figure 9: The 20 most frequently used TLMs. TLMs are counted once per study and grouped by family or model name when available (e.g., one BERT count for studies comparing BERT-base/large/(un)cased). Overall, the database includes 54 unique TLMs and 609 non-duplicate instances across studies, as illustrated above.

D.2 Additional Results to Linguistic Phenomena

Model type & language. Regarding coverage across TLMs, the distribution varies notably across syntactic phenomena, with the bidirectional TLMs dominating most phenomena (ranging from ∼60% to 73%) evaluation, while causal and both model types were more often tested on phenomena like *POS and other lexical properties* and *dependency parsing and constituency* (<20%). Regard

Fig. 19 for a distribution of studies across syntactic phenomena, model type, and interpretability method. When looking at the composition of phenomena within each model type, Fig. 10, *agreement and feature-checking* and *dependency parsing and constituency* constitutes the majority across all TLM tasks. Phenomena such as *argument structure and other constructions* is more prominent in causal TLMs.

Moreover, 62% of the articles focus on only one syntactic category, with 24% testing two, Fig. 11. This reflects a common emphasis on individual syntactic phenomena, which should not be interpreted as a limitation in itself. Rather, progress in the field benefits from combining fine-grained, phenomenon-specific analyses with more holistic approaches, as there is no single or definitive test for syntactic competence.

Across syntactic phenomena, monolingual TLMs dominate coverage in most categories (typically around ∼50–65%), while studies testing both and multilingual TLMs tend to contribute moderately (usually 15–30%) depending on the phenomenon, see Fig. 12.

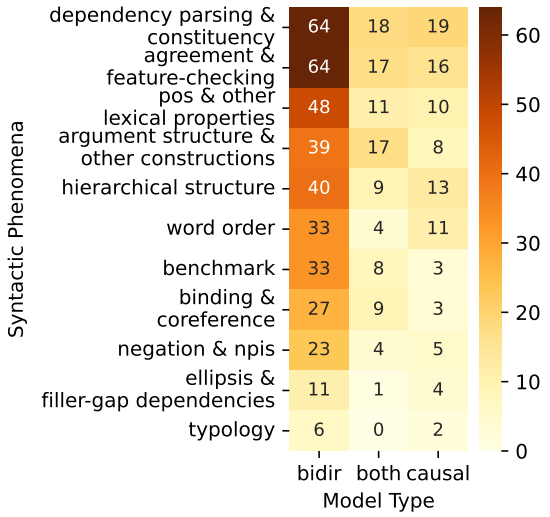


Figure 10: Heatmap of counts illustrating model type and syntactic phenomena.

D.3 Additional Results to Methods

Fig. 13 shows the counts of methods over the years, categorized according to whether they are observational or interventional. Notably, a significant temporal trend is that probing methods have diminished substantially in recent years, possibly due to the growing awareness of their limitations. For instance, following the publication of the influential criticism in Belinkov (2022), the number of probing studies with bidirectional TLMs in our database declined sharply, from 27 studies in 2022 to just 11 in 2023.

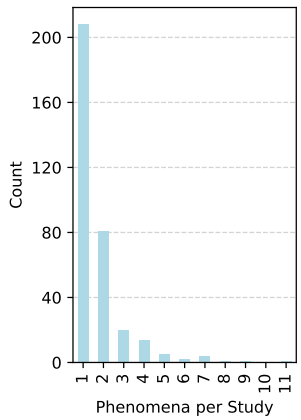


Figure 11: Counts of syntactic phenomena investigated per study.

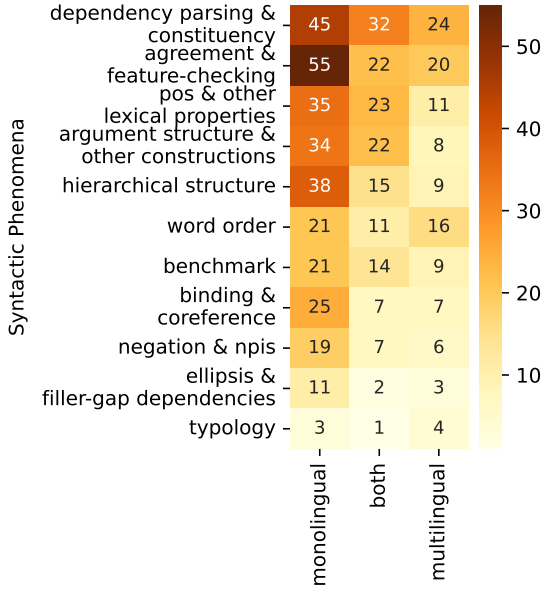


Figure 12: Counts heatmap of model language coverage and syntactic phenomena.

Linguistic phenomena. Fig. 14 shows a largely balanced distribution of methods across syntactic phenomena, with three notable exceptions. First, behavioral methods dominate for the phenomena that are most clearly at the syntax-semantics interface: *negation & NPIs* and *binding and coreference*. Second, mechanistic methods dominate in *benchmark*-based evaluations. Third, probing methods dominate in *dependency parsing & constituency* and *POS & other lexical properties*; for the latter, again we suspect this is an artifact of time, since these were among the first phenomena

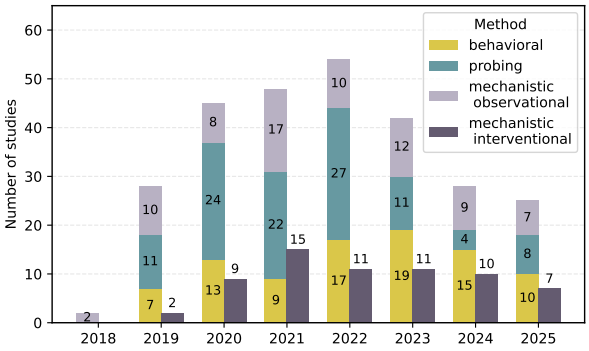


Figure 13: Counts of methods over years, grouped by observational and interventional approaches.

to be checked frequently with probing methods in earlier years. Moreover, tasks related to POS tagging and lexical properties are generally easier to formulate as classification problems than phenomena such as binding or argument structures, which may also contribute to this pattern.

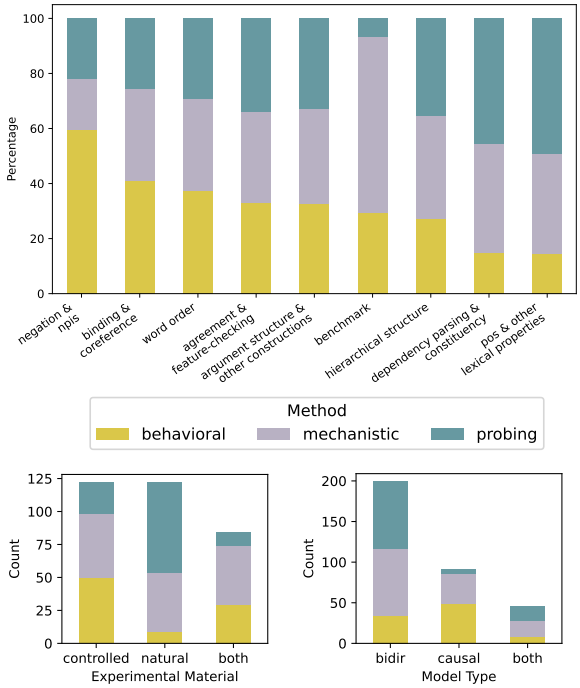


Figure 14: **Top:** Percentages of methods across syntactic phenomena; only phenomena with more than 20 instances are included. Linguistic phenomena are sorted by percentage of behavioral method. **Bottom:** Counts of experimental materials and model types by method (*unknown* experimental material with fewer than 10 counts excluded).

Experimental material. With respect to experimental materials, instead, clear differences emerge (Fig. 14, bottom left). Probing approaches (dark green) occur more frequently with naturalistic than with controlled data, particularly when targeting lexical or morphosyntactic properties such as POS (see Fig. 14, top); however, this pairing reflects an empirical co-occurrence rather than a methodological dependency, as the easy availability of e.g. POS annotations makes naturalistic data a convenient, but not necessary, choice for probing. Probing methods are instead almost absent in studies with causal TLMs (Fig. 14, bottom-right); we suspect that this is a historical accident, as most studies of causal TLMs have been carried out in 2024-25, when probing methods have become less common (note, however, that as mentioned above many mechanistic methods include some probing component).

Behavioral methods (lime yellow) pattern most frequently with controlled data, which makes sense given that many existing psycholinguistic datasets can be adapted to test TLMs, as behavioral methods need carefully designed data to test for a given property. Thus, the frequent pairing of behavioral methods with controlled data reflects the adaptation of psycholinguistic paradigms to test models, satisfying the need for carefully designed data to test specific properties.

E Additional Results to Section 4

E.1 Additional Results to BLiMP

Fig. 15 (top) shows the average BLiMP scores as a function of training data size, color-coded by model type. The first thing to note is how good models are according to this benchmark: the overall BLiMP score is 72% (random baseline: 50%). We also observe the well-known scaling pattern in deep learning: TLMs trained on larger amounts of data tend to perform better (linearly in log scale), as do larger TLMs (see also Fig. 15, bottom).

As for the difference between model types, note that, for a given size or amount of training data, bidirectional models tend to outperform causal models (as also observed in Waldis et al., 2024). Since BLiMP uses sentence-level comparisons, this cannot solely be due to the fact that bidirectional models, but not causal models, have access to the whole context of the stimulus when they process the critical fragment (e.g. *themselves/herself* in the first example in Fig. 7). We hypothesize instead that bidirectional models reach better sentence representations due to the bidirectional nature of the training task. However, for larger amounts of training data and model sizes

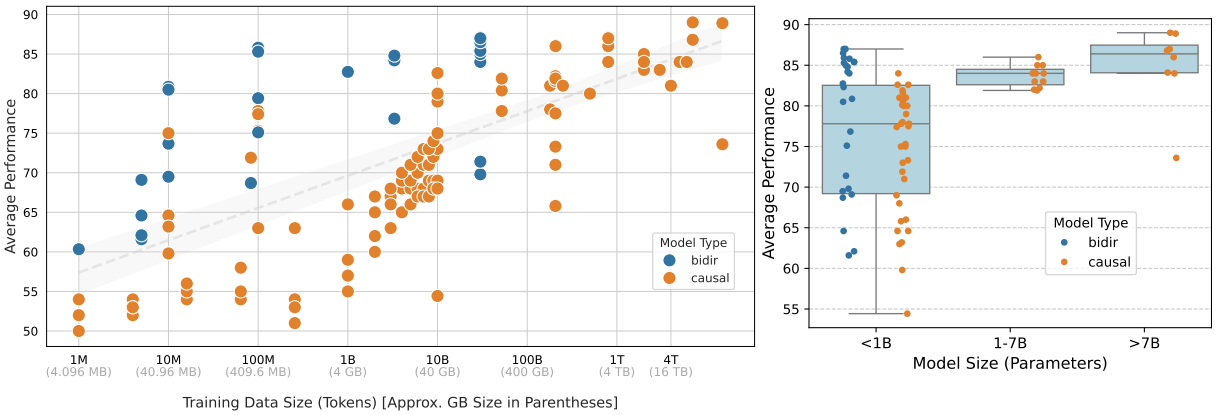


Figure 15: **Left:** BLiMP (Warstadt et al., 2020a) average scores plotted against the number of training tokens (log scale). **Right:** BLiMP average scores shown by model size and model type. If a paper reports scores for TLMs trained on varying sizes of training data, we include only the fully trained models in the boxplot.

	coeff.	SE
Intercept	77.57***	(1.85)
Model Type (causal)	-9.13***	(2.10)
Training Data Size (in billions)	0.0017***	(0.00)
Observations	134	
R^2	0.198	
Adj. R^2	0.186	
BIC	1003	
AIC	994	

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 4: OLS regression predicting average BLiMP accuracy scores. Standard errors in parentheses.

(>30B training tokens and >1B parameters) only causal models have been examined in the literature we cover, such that we cannot know whether the trend continues at these scales.

Statistical modeling. We fit an OLS regression model predicting average BLiMP accuracy scores, 134 data points in total. Full model specifications and estimates are reported in Table 4. Although the explained variance is modest ($R^2 = 0.186$, $AIC = 994.3$, $BIC = 1003$), both model type and training data size were significant predictors: causal models performed worse ($\beta = -9.13$, $SE = 2.095$, $p < 0.001$), and training data size showed a small positive effect ($\beta = 0.0017$, $SE = 0.0$, $p < 0.001$). The intercept ($\beta = 77.57$, $SE = 1.848$, $p < 0.001$) represents the expected score for a bidir TLMs at zero training data (represented in billions of tokens).

Variance across phenomena. We ranked TLMs by performance within each linguistic phenomenon and added slight jitter to handle ties. Then, we computed Kendall's τ correlations between these rankings across linguistic phenomena categories to see how similarly models perform on different linguistic phenomena. Fig. 16 presents a heatmap with the correlation score between each phenomenon. Kendall's τ correlations show strong and consistent relationships across most linguistic phenomena, with an average correlation of about 0.58 (SD $\approx$ 0.14), indicating very strong agreement in TLM performance across tasks. Quantifiers are a clear outlier, exhibiting weak to moderate correlations with other categories. In contrast, several pairs show especially high alignment, notably Argument Structure and Control/Raising (τ = 0.82) and Filler-Gap and Island Effects (τ = 0.75).

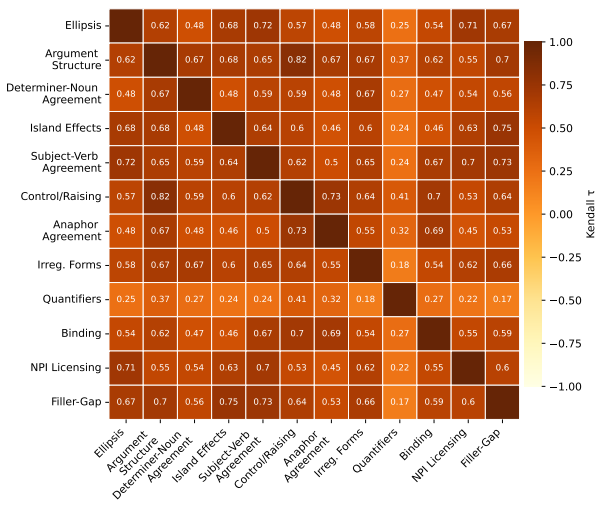


Figure 16: Computed model rankings per phenomenon, then Kendall's τ coefficient, visualized as a heatmap.

E.2 Additional Results to Beyond Benchmark and English

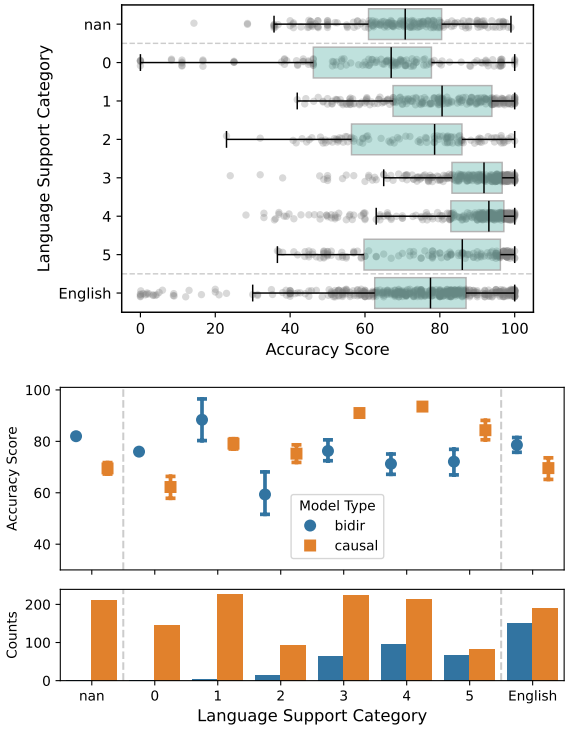


Figure 17: **Top:** Language support categories, categories taken from Joshi et al. (2020), with nan and English (category 5.) separately. **Bottom:** Mean accuracy scores by language group and model type.

Fig. 17 displays the distribution of TLM accuracy scores grouped by language support, based on the five categories reported by Joshi et al. (2020) and model type (bottom).

Statistical modeling. We estimate an OLS regression model predicting accuracy scores as a function of model type, linguistic phenomenon, and language support category. To ensure comparability across languages, we exclude instances with missing language support annotations (i.e., "nan" category in Fig. 17), as well as all scores annotated as *benchmark*. The resulting analysis is performed on 1,577 observations. Language support is treated as an ordered categorical variable (following Joshi et al. (2020)). Full model specifications and estimates are reported in Table 5, and a forest plot of the estimated coefficients is shown in Fig. 18.

E.3 Additional Results to Probing

We analyze 147 studies in total that employ probing methods, which includes all papers coded as probing, plus some coded as mechanistic when they also employed probing methods. Fig. 20 shows that probing-based studies successfully identify syntactic information at comparable rates across linguistic phenomena categories.

	coeff.	SE
Intercept	0.43***	(0.02)
Model Type (causal)	0.099***	(0.01)
Language support	0.05***	(0.003)
Agreement & feature-checking	0.17***	(0.01)
Argument structure	-0.17***	(0.02)
Binding & coreference	0.06	(0.04)
Dependency parsing	-0.003	(0.02)
Ellipsis & filler–gap depend.	-0.06	(0.04)
Hierarchical structure	-0.09***	(0.02)
Negation & NPIs	-0.11**	(0.03)
POS & other lexical properties	0.19***	(0.03)
Observations	1577	
R^2	0.23	
Adj. R^2	0.22	
BIC	-784.1	
AIC	-843.1	

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 5: OLS regression predicting accuracy scores. Standard errors in parentheses.

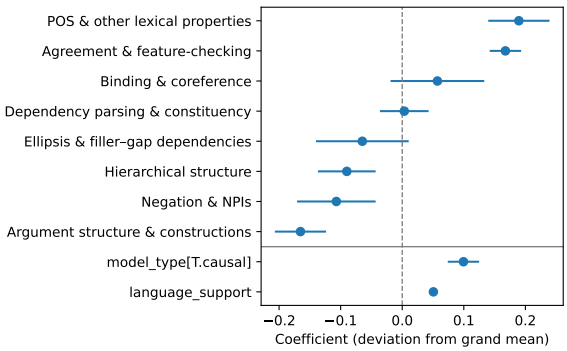


Figure 18: Forest plot of OLS regression coefficients predicting accuracy scores. The model includes model type, linguistic phenomenon (except *benchmark*), and language support category. Points represent coefficient estimates and horizontal lines indicate 95% confidence intervals.

E.4 Additional Results to Mechanistic Interpretability

We aim to characterize the mechanistic interpretability field by categorizing all 140 mechanistic papers by level of analysis and method. To abstract over this diverse set of approaches and assess what TLMs learn about syntax, we evaluate whether studies report evidence of syntactic knowledge or instead highlight its limitations.

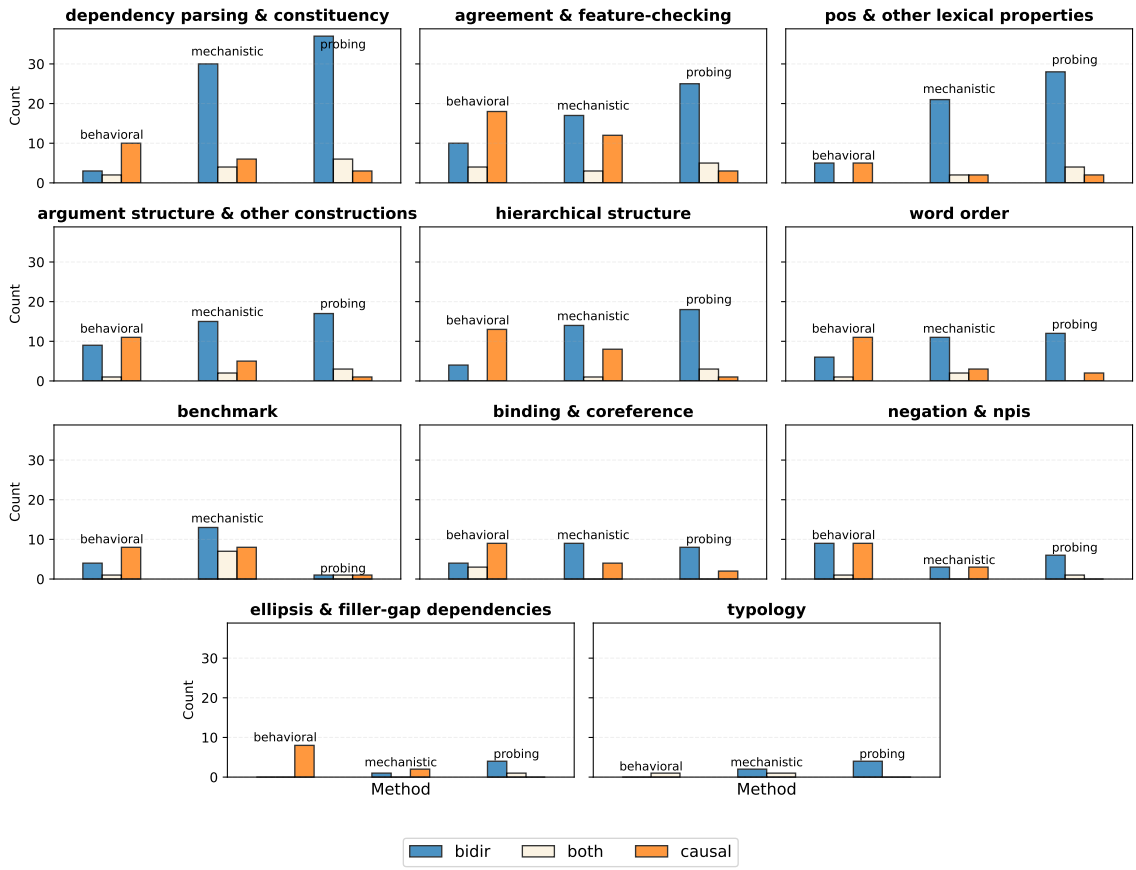


Figure 19: Distribution of studies in our database across syntactic phenomena, model type (bidirectional, causal, both TLMs types) and interpretability method, ordered by overall counts.

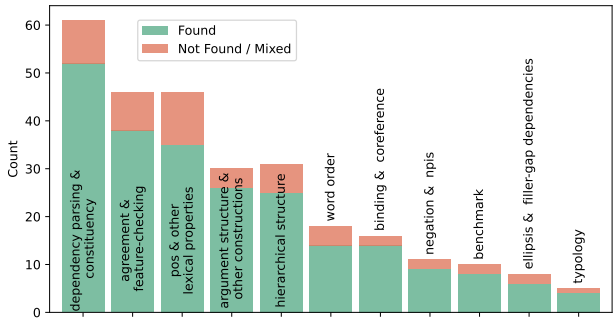


Figure 20: Successful recovery of syntactic information from TLMs' representations via probing across linguistic phenomena.

Overall few in counts, the mixed and critical findings suggest that models rely heavily on surface-level patterns rather than abstract syntactic rules, exhibiting limited and inconsistent sensitivity to structural changes. Also, TLMs' syntactic abilities vary across conditions, improving with more data and in high-resource languages but remaining weak in low-resource settings, and can even degrade with fine-tuning.

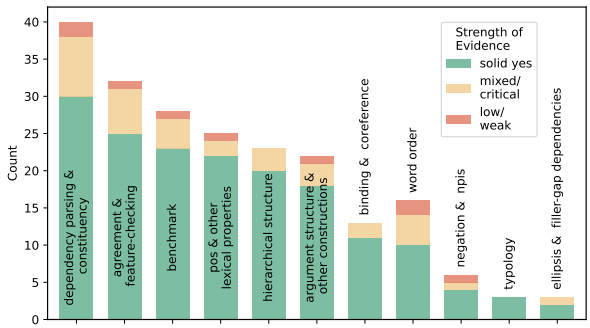
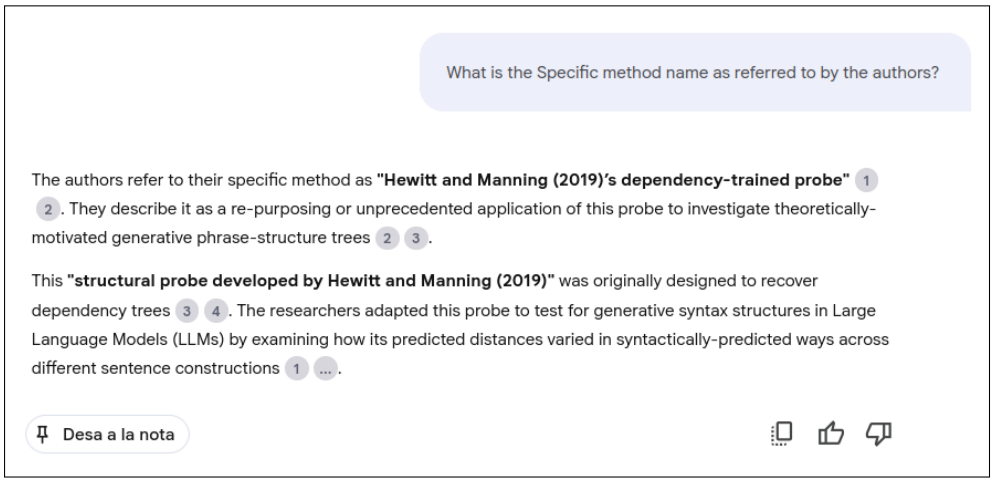


Figure 21: Evaluation of syntactic awareness in mechanistic interpretability studies by linguistic phenomena tested.



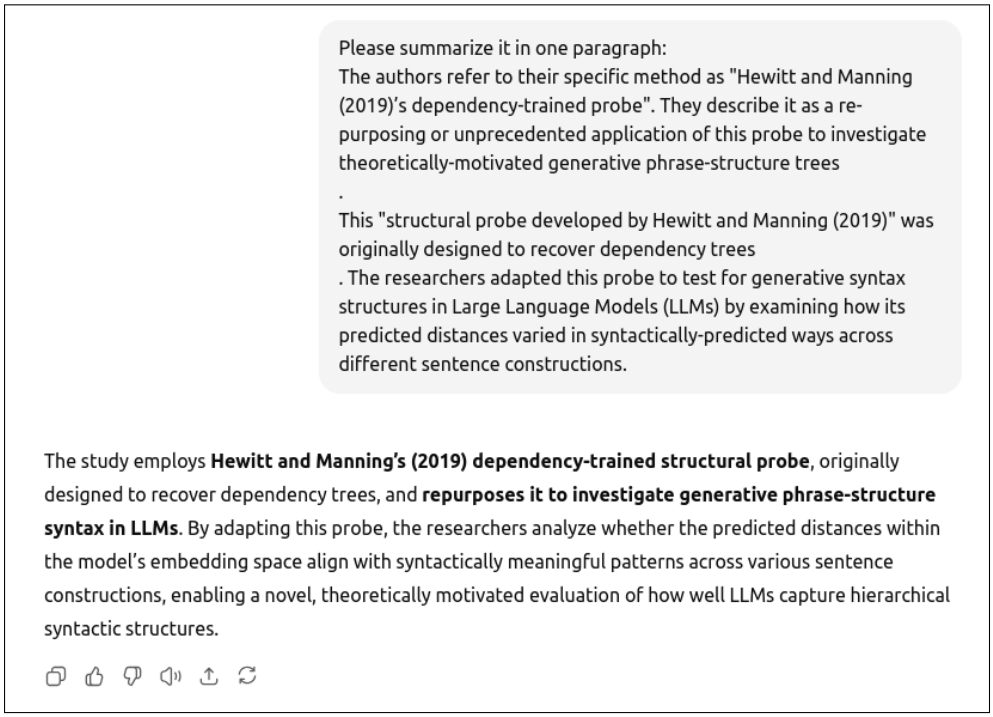


Figure 22: Illustration of the AI-based workflow. **Top:** NotebookLLM (Google Labs, 2023) interaction depicting how the model assists with information extraction; numbers refer to the original source in the paper. **Bottom:** Example of a ChatGPT (OpenAI, 2025) session showing a query and the generated response.

You are evaluating textual information. Given a paragraph and a note, which are based on the same source, rate how much the note captures the information of the paragraph on a scale from 0 to 5, where 0 = unrelated and 5 = fully equivalent.

1 majority of the information about the syntax tree of a sentence is in fact extractable by the model. This, however, is not true for randomly initialised trans- former models. results, thus, lead us to conclude that a transformer's training leads its attention heads to have the potential to decode syntax trees.

#2 demonstrates that transformer-based language models (LMs) like BERT, ALBERT, and RoBERTa have a high potential to extract syntactic information from sentences, as measured by the Architectural Bottleneck Principle and attentional probes. Using V-information, the study found that these models could access roughly 85% of syntactic information in English sentences, with similar trends across other languages, and that this information is most extractable in the middle layers. Importantly, this syntactic knowledge emerges from pretraining rather than being inherent to the architecture. However, while attention heads could use this information, the study shows they do not actually compute weights resembling syntax trees, leaving the question of actual usage open. Methodologically, the study also highlights that simpler structural probes underestimate LMs' potential, as the attentional probe recovered more syntactic information than previous baseline methods.

I would rate the note's coverage of the paragraph as **3 out of 5**.

Figure 23: Instance of a ChatGPT (OpenAI, 2025) prompt showing our evaluation query with an example.